The Impact of Epigenetic Factors on Nucleotide Substitution Patterns in Mammalian Genomes

By

Hua YING

A thesis submitted for the degree of Doctor of Philosophy
of The Australian National University

June 2010

Declaration

This thesis contains no material which has been accepted for the award of any other degree. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due acknowledgement has been made.

Name: Hua Ying

Signature: .....

Date: 21/06/2010.

Acknowledgements

I would like to express my sincere gratitude to Associate Professor Gavin Huttley for his thorough and inspiring supervision. His enthusiasm for science and broad knowledge has set an example for me and motivated me to pursue my academic career. I am certainly looking forward to opportunities to work with, and learn from him in the future.

I am also grateful to the members of my advisory panel, Dr Rohan Williams, Professor Simon Easteal, and Professor David Tremethick. They have provided invaluable support and advice at various stages.

I would particularly like to acknowledge my collaborator, Dr Julien Epps, from UNSW for his valuable input to the signal processing analyses. It was a very rewarding and enjoyable collaboration which was certainly the highlight of my project. In addition, I would like to thank Dr James Cai from Stanford University for assistance in getting yeast data.

I would like to thank all members of the computational genomics group and my friends at JCSMR for their help and in providing a pleasurable environment to conduct my research. Especially, I thank Helen Lindsay for valuable discussions and in sharing her experience in using LaTeX.

I greatly appreciate my husband Xun Han, my daughter Angela Han and my parents for their unwavering support and love throughout my research. Their understanding has been a source of strength for me to complete the thesis.

Finally, I am grateful to Dr Stephen Ohms for his careful proofreading.

Abstract

Epigenetic factors serve as a bridge that connects genetic information and cellular activities. These effects are achieved through chemical modifications to DNA or the histone proteins that modify DNA structure. Dynamic changes in epigenetic marks flexibly regulate outcomes of DNA functions. Despite this essential responsibility in eukaryotic biology, epigenetic factors have been implicated in profoundly impacting on mutagenesis which may affect DNA sequence evolution. That epigenetic factors affect both lesion formation and DNA repair has been firmly established by experimental evidence, but their influences in substitution have remained elusive.

In this thesis, I examine the relationship between epigenetic factors and sequence substitutions. Investigated epigenetic factors are chromatin structure and DNA methylation. Depending on their size distribution and influences on mutagenesis, these factors lead to substitution rate heterogeneity at different scales with distinct attributes.

First order chromatin compaction imposes a physical barrier to mutagenesis and causes substitution rate and type heterogeneity. I first sampled DNase I hypersensitive sites (DHS) and their flanking sites (Flank) as representative of a relatively open and closed chromatin structure respectively. My analyses revealed that both total substitution and transition substitution rates were lower at DHS than Flank. Since the major difference in chromatin between DHS and Flank is due to nucleosome organization, I then evaluated the influence of individual nucleosome

positioning on substitution. The distribution of substitution rate was found to oscillate along the promoter sequence with a dominant periodicity of ~200 bp. A comparable oscillation was detected from experimental measurements of nucleosome density signals. These observations strongly support a contribution from nucleosome placement to localised substitution rate heterogeneity.

The modified base 5-methylcytosine (5^mC), which is confined to the CpG dinucleotide in vertebrates, greatly influences protein-coding sequence evolution under the joint effects of mutation pressure and natural selection. 5^mC exhibits a high mutation rate due to an increased spontaneous deamination rate, but selective constraints on functional 5^mC can oppose this mutation pressure. This phenomenon predicts that in methylated genomes a distinctive mutation-selection balance will result in stark differences in CpG equilibrium frequency and evolutionary rate between positions that differ in mode of selection. Examining substitution rate within a CpG context demonstrated that the rate of CpG transitions was highly elevated for the majority of primate genes, but only slightly elevated for a minority of genes in yeast which lack DNA methylation. Selective constraints were also stronger on CpG codons than other codons in primates with the dominant effect being purifying selection. Furthermore, genes with stronger natural selection on CpG codons were enriched in disease-associated genes. These results suggested that CpG codons occupy functionally more important positions in primate genes.

In conclusion, my work demonstrates that epigenetic factors have a profound affect on the distribution of genetic variation. Experimentally determined effects on mutation processes determined do correspond with substitution rate changes. A corollary to the relationship between epigenetic state and genetic variation is that the resulting impression in DNA sequence evolution may potentially be applied to improved understanding of the organization and function of epigenetic factors.

Contents

Declaration i

Acknowledgements iii

Abstract v

Table of Contents vii

List of Figures xiii

List of Tables xv

Abbreviations xvii

Objective and Roadmap 1

List of Publications 5

1 Introduction 7

1.1 Evolutionary rate heterogeneity 7

1.1.1 Large scale variation 8

1.1.2 Localized variation 9

1.1.3 Among sites variation 9

1.1.4 Variation in the profile of sequence divergence 10

1.2 Mechanisms of mutagenesis 10

1.2.1 Processes of lesion formation 11
1.2.2 Processes of lesion repair 12
1.3 Causes of evolutionary rate heterogeneity 13
1.3.1 Natural selection 13
1.3.2 Epigenetic factors 14
1.3.3 Other biological factors 16
1.4 Evolutionary rate measurement 17
1.5 A maximum likelihood approach to measure the substitution rate 18
1.5.1 Substitution models 18
1.5.2 Maximum likelihood estimation 20
1.5.3 Hypothesis testing 22

2 Programmatic Sampling of Genomic Data 25
2.1 Motivation 25
2.2 Features 29
2.2.1 Design 29
2.2.2 Ensembl database schema 29
2.2.3 Implementation 33
2.2.4 Key modules 33
2.3 Examples 39
2.3.1 Example 1: getting a sequence with annotations for a genomic coordinate 39
2.3.2 Example 2: retrieving all human protein-coding genes and their coding sequences from canonical transcripts 40
2.3.3 Example 3: retrieving gene sequence alignments on human chromosome 1 for human, macaque and chimpanzee 41
2.3.4 Example 4: retrieving human and mouse ortholog protein-coding genes with a one-to-one relationship 43
2.4 Conclusion 44
2.5 Requirements 45

3 Evidence that Chromatin Compaction Affects Substitution Rate 47

3.1 Motivation 48

3.2 Methods 52

3.2.1 Data 52

3.2.2 Software 54

3.2.3 Statistics 54

3.3 Results 57

3.3.1 DHS regions exhibit a distinct substitution rate 57

3.3.2 DHS regions exhibit a distinct substitution type profile . . 58

3.3.3 Intergenic, but not intronic, regions exhibit a distinct CpG transition rate 60

3.3.4 Substitutions resulting from CpG methylation do not completely account for differences in transition substitutions between DHS and Flank regions 61

3.3.5 Purifying natural selection on functional elements does not appear to be a cause of substitution heterogeneity 61

3.4 Discussion 64

4 Evidence That Nucleosome Placement Contributes to Localised Substitution Rate Heterogeneity 67

4.1 Motivation 68

4.2 Methods 72

4.2.1 Promoter data with nucleosome annotations 72

4.2.2 Chip-seq nucleosome signals 73

4.2.3 A Phylogenetic hidden Markov model to measure spatial substitution rate heterogeneity 73

4.2.4 Phylogenetic footprinting to measure spatial substitution rate heterogeneity 75

4.2.5 Statistical testing of the correlation between substitution spectrum and nucleosome score 76

4.2.6 Signal period estimation 76
4.3 Results 78
4.3.1 The substitution rate was significantly heterogeneous along promoter sequences 78
4.3.2 Phylo-footprinting displayed similar substitution rate heterogeneity 79
4.3.3 The spatial substitution spectra and nucleosome scores are significantly correlated for some loci 81
4.3.4 An $\sim$ 200 bp oscillation in both substitution rate and nucleosome score 87
4.4 Discussion 90

5 The Impact of DNA Methylation on Protein Coding Sequence Evolution **97**
5.1 Motivation 98
5.2 Material and Methods 102
5.2.1 Statistical models of codon evolution 102
5.2.2 Hypothesis Testing 108
5.2.3 Data sampling 111
5.3 Results 114
5.3.1 Elevated CpG transition and transversion rate were evident 117
5.3.2 CpG transitions were the major context-dependent effect . 120
5.3.3 Methylation-affected amino acids exhibited a different non-synonymous substitution rate 122
5.3.4 CpG-encoded amino acids were subjected to stronger purifying selection in primates than in yeast 123
5.3.5 Genes displaying significant CpG effect were enriched in disease-causing genes 125
5.3.6 Within-species genetic variation analyses further supported purifying selection affecting exonic CpG polymorphism . . 127

5.4 Discussion 128
5.5 Supplementary 134
5.5.1 Modeling based on the Y98 model 134
5.5.2 Assessment of statistical power 137

Conclusion **139**

Bibliography **143**

List of Figures

1.1 **Q** matrices for the F81 and HKY models 19
1.2 Unrooted phylogenetic tree for four species a-d 21

2.1 Ensembl variation database schema 31
2.2 Ensembl core database schema for gene-related tables 32

4.1 Comparison of K from phylo-HMM and footprinting 80
4.2 Comparison of K with nucleosome annotations 82
4.3 Q-Q plot of the bootstrap test p distribution against the uniform distribution 85
4.4 Comparison of K with nucleosome scores 86
4.5 Signal analysis of K by DFTs 88
4.6 Periodicity estimated from K and nucleosome scores with CRB $<$ 0.20 89
4.7 Periodicity estimated from K and nucleosome scores with various CRB thresholds 95

5.1 Flow diagram for nested model parameterization and LRTs 110
5.2 Phylogenetic tree for primates and yeast 111
5.3 $\hat{G}$ distribution estimated from CNF+G model with $p_1 < 0.05$ . . . 119
5.4 $\hat{G.K}$ distribution estimated from CNF+G.K model with $p_3 < 0.05$ 120
5.5 $\hat{G}$ distribution estimated from CNF+G+G.K model with $\arg\max(p_3, p_4) <$ 0.05 121

5.6 $\hat{G.K}$ distribution estimated from CNF+G+G.K model with $\arg\max(p_1, p_2) < 0.05$ 122
5.7 $\hat{\alpha}$ distribution estimated from CNF+G.K+α model with $\arg\max(p_3, p_5) < 0.05$ 123
5.8 $\hat{\alpha}$ distribution estimated from CNF+G.K+G.K.$\omega + \alpha$ model with $\arg\max(p_3, p_7, p_8) < 0.05$ 124
5.9 $G.\hat{K}.\omega$ distribution estimated from CNF+G.K+G.K.$\omega + \alpha$ model with $\arg\max(p_3, p_5, p_6) < 0.05$ 125
5.10 $G.\hat{K}.\omega$ distribution estimated from CNF+G.K+G.K.ω model with $\arg\max(p_3, p_7) < 0.05$ 126
5.11 Statistical power of LRTs 138

List of Tables

3.1 Support for substitution rate and type differences between DHSs and Flanks 59

3.2 Support for substitution rate and type differences between DHSs and Flanks after eliminating alignments containing conserved elements 63

4.1 Promoters with correlated K and nucleosome scores 84

5.1 Substitution model terms 105

5.2 Codon substitutions represented by α and $G.K.\omega$. 107

5.3 Statistics from analyses of primate *BRCA1* using the CNF baseline model 115

5.4 Statistics from analyses of primate *F8* using the CNF baseline model 116

5.5 Number of significant genes from paths I and II 118

5.6 Number of significant genes from paths III and IV 118

5.7 Testing for enrichment of CpG-affected genes in OMIM disease-causing genes 127

5.8 Classification of human biallelic SNPs within coding regions . . . 129

5.9 Statistics from analyses of primate *BRCA1* using the Y98 baseline model 135

5.10 Statistics from analyses of primate *F8* using the Y98 baseline model 136

Abbreviations

5^mC	5-methylcytosine
BER	Base excision repair
CRB	Cramer-Rao bound
DFT	Discrete Fourier transform
DHS	DNase I hypersensitive site
Flank	Flanking site of DHS
GC	Percentage of G+C in the sequence
HMM	Hidden Markov model
LRT	Likelihood ratio test
OMIM	Online Mendelian Inheritance in Man
NER	Nucleotide excision repair
SGD	Saccharomyces Genome Database
SNP	Single nucleotide polymorphism
SQL	Structured query language
TCR	Transcription-coupled repair
TFBS	Transcription factor binding site

TSS	Transcription start site
UTR	Untranslated region

Objective and Roadmap

The objective of this thesis is to elucidate the impact of epigenetic factors on evolutionary divergence of primate genomes. Examples of such epigenetic factors include chromatin structure and DNA methylation. Depending on the magnitude of the epigenetic factors, their influence occurs on different scales of genomic sequences. For example, the basic chromatin packaging units, the nucleosomes, appear to be responsible for an oscillation in evolutionary rate with a period of ~200 bp, while hypermutable methyl-CpG sites influence substitution rates at a single nucleotide level. Consequently, in this thesis, a broad range of phylogenetic models and methods were applied to address questions including: (1) does evolutionary rate heterogeneity at different genomic sequence scales correspond to different epigenetic factors? (2) how do epigenetic factors affect genetic variation? (3) does the influence of epigenetic state arise from an influence on lesion formation, DNA repair and/or natural selection? (4) what impact does the heterogeneity in genetic variation resulting from epigenetic factors have on genetic encoding of phenotype? These questions are evaluated in the following chapters.

Chapter One gives an introduction to the concept of rate heterogeneity on different scales, mechanisms of mutagenesis, and proposed causes of evolutionary rate heterogeneity. Additionally, because the major techniques applied in the thesis are in the realm of comparative genomics, methods to measure evolutionary rate from comparative data and maximum likelihood-based phylogenetic approaches

are introduced.

Chapter Two describes an Ensembl-querying module developed for inclusion in the PyCogent library. It provides capabilities for querying and retrieving various types of genomic data from Ensembl databases. In this chapter, the schemas of the Ensembl databases are introduced. Key modules and major capabilities are described. Some examples of PyCogent code used to carry out genomic data sampling, similar to those in the following chapters, are explained in detail.

Chapter Three addresses issues of differences in the rates and types of substitution between open and closed chromatin structures. Open chromatin structures are DNase I hypersensitive sites (DHSs), whereas their flanking regions (Flanks) contain closed chromatin structures. From likelihood ratio tests, questions addressed were whether open chromatin shows different rates and types of substitution from those of closed chromatin. In addition, given the connection between CpG methylation and chromatin compaction, the question of whether CpG sites contribute to differences in the types of substitution between open and closed chromatin was examined.

Chapter Four presents evidence that nucleosome placement affects the spatial distribution of substitution rate, which is defined as the **substitution spectrum**. Substitution rate heterogeneity in promoters was evaluated by phylogenetic footprinting and phylogenetic hidden Markov models. The resulting substitution spectra were compared with experimentally defined nucleosome density signals. In addition, patterns of substitution spectra were estimated using signal processing techniques and compared with those estimated from nucleosome mapping signals.

Chapter Five addresses the impact of CpG methylation on the evolution of protein-coding sequences. Such an impact is represented by a shifted mutation-selection balance associated with hypermutable 5^mC nucleotides located in coding sequences. This property was compared between primates whose genomes are

heavily methylated and yeast whose genomes are putatively free of methylation. Through the use of appropriate codon substitution models and context-dependent parameters, substitution rates and natural selection on CpG sites were evaluated and compared for the two clades. As variation in selective constraints will affect phenotype, the question of whether genes that exhibit stronger natural selection operating on CpG codons than other codons are enriched in disease-causing genes was investigated using data from the literature. Additionally, CpG-mutation and selection properties were further evaluated in a genome-wide coding SNP survey in human.

The concluding chapter summarises my findings, their implications, and discusses potential future work emanating from this thesis.

List of Publications

The following publications are based on work carried out during my period of PhD candidature.

1. Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC, Eaton M, Hamady M, Lindsay H, Liu Z, Lozupone C, McDonald D, Robeson M, Sammut R, Smit S, Wakefield MJ, Widmann J, Wikman S, Wilson S, **Ying H**, Huttley GA: **PyCogent: a toolkit for making sense from sequence.** *Genome Biol* 2007, **8**(8):R171

   URL: `http://genomebiology.com/2007/8/8/R171`

   This paper presented PyCogent version 1.0. Chapter Two describes an ensembl module that is one of the recent additions to PyCogent.

2. Lindsay H, Yap VB, **Ying H**, Huttley GA: **Pitfalls of the most commonly used models of context dependent substitution.** *Biol Direct* 2009, **4**:10

   URL: `http://www.biology-direct.com/content/4/1/10`

3. **Ying H**, Epps J, Williams R, Huttley GA: **Evidence that localised variation in primate sequence divergence arises from an influence of nucleosome placement on DNA repair.** *Mol Biol Evol*: In press

   URL: `http://mbe.oxfordjournals.org/cgi/content/abstract/msp253?ijkey=Ml5IiBBWVNyhdm0&keytype=ref`

   Chapter Three and Four describe work in this paper with further details.

Chapter 1

Introduction

Abstract

In this chapter, I introduce the concept of evolutionary rate variation, along with the evidence that epigenetic factors contribute to it. Variation in the rate of sequence divergence is termed rate heterogeneity, and may exist on scales between those of immediate nucleotide neighbors to entire chromosomes. Many biological factors that affect mutagenesis have been proposed to explain this phenomenon. Among these factors, special attention is given to the potential contribution of epigenetic factors, although their impact remains elusive. I then present the measurement techniques that will be used in this thesis to evaluate the effects of epigenetic factors on evolutionary rate heterogeneity. In particular, maximum likelihood-based phylogenetic models and their applications in hypothesis testing are introduced.

1.1 Evolutionary rate heterogeneity

Rates of evolution are not constant within a genome and are distributed on various scales. They are observed at chromosome levels, regional levels ranging from

hundreds of base pairs to millions of bases pairs, and localized levels as small as single nucleotides. Additionally, the profile of evolutionary divergence also varies substantially across genomes.

1.1.1 Large scale variation

A traditional view of evolutionary rate heterogeneity is on a scale from millions of base pairs to entire chromosomes. In mammals, a general observation is that evolutionary rates are highest on the Y chromosome, followed by autosomes, and lowest on the X chromosome [1, 2, 3]. Among autosomes, heterogeneity is also marked, in that the evolutionary rate of eight of the autosomes is significantly different from the mean human autosomal rate of evolution [3]. Furthermore, within chromosomes, substantial variations have been demonstrated at the 1Mb level for both genic [3] and intergenic [4, 5] sequence divergence.

Another important aspect of evolutionary rate heterogeneity is its covariance with certain chromosomal features, such as background GC content and chromatin status. Nucleotide composition in vertebrate genomes is characterized by "isochores" that are long regions ($>$ 300 kb) of relatively homogenous GC content. Analyses of synonymous substitution rates [6] and polymorphisms [7] have revealed a positive correlation between substitution rates and GC content. A recent examination has established the presence of evolutionary rate heterogeneity between open and closed chromatin [8], which was defined from chromatin fibres of approximately 1Mb in length in the human genome. The evolutionary rate was found to be significantly negatively correlated with the openness of chromatin structure in intergenic regions, ancient repeats, and introns.

1.1.2 Localized variation

Local scale evolutionary rate heterogeneity ranges over distances from hundreds to thousands of base pairs and is widely spread throughout genomes. For example, functional sequence components mostly fall into this category and their observed evolutionary rates follow the order: coding sequences < 5' or 3' UTRs < introns / promoters [9]. Intergenic regions also harbor conserved sequences that are distinguishable from the background rate [10].

Apart from these known functional categories, however, local variation in evolutionary rates may be potentially more universal than is currently accepted. For instance, evolutionary rates differ between nucleosomes, which are the basic chromosome packaging units, and linkers which are the DNA sites between nucleosomes [11, 12, 13, 14]. Another study [15] in fish genomes has revealed periodic changes in evolutionary rates downstream from transcription start sites (TSSs) corresponding to nucleosome positioning. Although all these analyses were based on concatenated sequences, they have established a correlation between nucleosome placement and evolutionary rate. Because nucleosomes occur repeatedly along the chromosome, they are likely to generate widespread local rate heterogeneity.

1.1.3 Among sites variation

The evolutionary rate is influenced by neighboring sites and differs from site to site. This effect has been known for some time and observed in both experimental assays of DNA replication [16] and analyses of evolutionary rates [17]. One of the most prominent sequence context effects occurs at CpG sites arising from 5-cytosine methylation (5^mC) in vertebrate genomes. Due to their high rates of deamination [18], methyl-CpG sites exhibit a much higher spontaneous mutation rate than that of other dinucleotides [17, 19, 20, 21]. Moreover, codons

are a naturally-occurring context-dependent trinucleotide unit due to their explicit coding function. A common observation is that the mutation rate at codon position III is higher than at the other two codon positions because of relaxed selective constraints at synonymous sites.

1.1.4 Variation in the profile of sequence divergence

In addition to the overall evolutionary rate, the types of evolutionary changes vary greatly across genomes. Substitutions can be classified into two major categories: transitions and transversions. Transitions are substitution mutations from purine to purine or from pyrimidine to pyrimidine, while transversions are interchanges between purine and pyrimidine. Although there are four possible transitions and eight possible transversions, transitions occur more frequently than transversions. The ratio of transitions to transversions, which I will refer to as λ, is expected to be greater than 1 [22, 23, 17], and is heterogeneous across the genome. The mean λ has been estimated to be 4.26 with a variance of 9.65 from mouse-rat gene comparisons [24]. Large variances in λ still exist after removal of hypermutable CpG sites. Moreover, closely located genes show more similar levels of λ compared to genes on different chromosomes [24]. This positive correlation between physical proximity and both the rates and types of substitution suggests they may originate from a common cause.

1.2 Mechanisms of mutagenesis

The multiple scales of heterogeneity in evolutionary rate indicate that they may have different origins. Thus, understanding the mechanisms that affect DNA mutations in the cell is essential because these underlie all genetic variations. Formation of lesions in DNA and their repair are the two key elements to mutagenesis: lesions create defects in genomic DNA, whereas DNA repair systems

target these defects to repair them and maintain DNA sequence integrity. There are many different mechanisms by which DNA lesions can be formed and, in turn, be repaired.

1.2.1 Processes of lesion formation

Lesions may occur spontaneously in DNA due to chemical reactions at various stages of the cell cycle or be induced by environmental factors such as ultraviolet light (UV). Depending on the cause of the lesion, different types of mutation may result.

Errors in DNA replication are one source of mutation. They can occur when bases in nucleotides change from their usual stable forms to temporarily less stable forms. Such temporary changes are called tautomeric shifts and can alter nucleotide pairing properties and lead to the formation of unusual base pairs, e.g. G-T, A-C, G-A pairs. As a result, spontaneous tautomeric shifts cause transitions or transversions during DNA replication [23].

Base modifications are another source of mutation. For instance, cytosine deamination removes the amino group on the number 2 carbon of C and converts it to U. If cytosine is methylated at the fifth position of the pyrimidine ring, the product of cytosine deamination becomes T. In both situations, this produces transition mutations.

Physically-induced lesions are another common type of DNA defect. For example, exposure of cells to UV-irradiation results in several types of mutagenic photoproduct, like cyclobutane pyrimidine dimers and (6-4) photoproducts [25, 26]. The major mutational events from UV-damage are G:C $\rightarrow$ A:T transitions [27, 28].

1.2.2 Processes of lesion repair

While lesions in DNA are common, DNA repair acts to maintain a low mutation rate. Some mechanisms of DNA repair include:

DNA mismatch repair (for a review, see [29]) corrects mismatched nucleotides arising from replication errors, recombination and several classes of DNA damage. During DNA replication, mismatch repair specifically recognizes and repairs errors in the newly synthesized strand that have escaped correction by proofreading. In vertebrates whose genomes are methylated, mismatch repair is responsible for correcting T-G mispairs arising from 5^mC deamination [30].

Excision repair includes a variety of activities that correct different types of DNA lesions. The common features of the excision repair pathways include recognition of the lesion site, removal of the lesion and possibly some flanking DNA bases from the damaged strand, DNA synthesis by a DNA polymerase to fill the gap, and finally restoration of a functional chromatin structure. There are two major types of excision repair, base excision repair (BER) and nucleotide excision repair (NER) (for a review, see [31]). They differ in that BER only removes the damaged site while NER recognizes bulky distortion of the DNA helix and removes a short segment containing the lesion [31, 32]. They are responsible for repairing a variety of types of damage, including photoreactivation products from UV damage [27, 26], and cytosine deamination resulting U-G pairs [33].

Of particular significance is transcription-coupled repair (TCR), a repair pathway operating specifically on the transcribed strand [34, 35]. When RNA polymerase II encounters a lesion in DNA during transcription, it stalls without further elongation. In eukaryotes, a few nucleotides at the 3' end of the nascent RNA are then removed, followed by excision repairs (BER or NER). Once the repair is complete, RNA polymerase II continues transcription. Thus, DNA repair is more frequent and more rapid at actively transcribed genes.

1.3 Causes of evolutionary rate heterogeneity

The causes of evolutionary rate heterogeneity can be classified as either selection-driven or mutation-driven. Natural selection operates on genomic sequences that encode information affecting phenotype such as exons, RNAs, and regulatory elements. Other factors that contribute to, produce, or reduce mutations are also candidate causes of variation in sequence divergence rate. These include many biological factors that affect mutagenesis, such as the number of cell divisions, replication errors, biased DNA repair and, as is now emerging, epigenetic factors. Each factor affects sequence divergence in a distinct manner leading to different patterns of evolutionary rate heterogeneity.

1.3.1 Natural selection

Natural selection operates on the genetic variants generated by the mutagenic processes described above. The influences can be suppressive or accelerative depending on the phenotypic effect of the mutation products. For instance, when a deleterious mutation that significantly reduces the fitness of an individual occurs, natural selection will oppose this mutation by preventing it from becoming fixed in the population. Such a constraint is termed *purifying selection* and regions, e.g most exons, subjected to strong purifying selection exhibit a low rate of evolution. This effect can be directly observed from the "conservation" tracks in the UCSC genome browser, because sharp transitions occur in conservation scores at the boundaries of exons and introns. Additionally, the highly conserved non-coding regions that exhibit a similar conservation level to exons are putative regulatory domains [10, 36, 37].

Natural selection, however, cannot account for all evolutionary rate heterogeneity, especially in regions where a functional role is ambiguous. It has been estimated that ~3-8% of the DNA in the human genome is subjected to natural selection

[10, 38, 39]. Thus, the majority of intronic and intergenic sequences are evolving almost neutrally. The observed widespread evolutionary rate heterogeneity in these regions, such as evolutionary heterogeneity among different chromatin states in ancient repeats, cannot be explained by selective constraints. Moreover, the causes of the periodic patterns in evolutionary rate downstream of TSSs in fish genomes [15] appeared to be unrelated to natural selection, even though the sequences themselves were under the influence of natural selection.

1.3.2 Epigenetic factors

Epigenetic factors are heritable factors that affect the development and function of a cell or organism without changing its DNA sequence. Such factors include chromatin structure, DNA methylation, and histone modifications. Their existence alters DNA susceptibility to mutagenesis.

Chromatin affects both lesion formation and repair efficiency through its 3-dimensional structure. In vivo, a DNA strand wraps around a histone octamer comprising eight histone proteins to form a DNA-protein unit of 147 bps called a nucleosome. Upon the binding of histone proteins, nucleosome-associated DNA sites may hinder the binding of the repair machinery, although they are also relatively protected from lesion formation. Conversely, the nucleosome-free sites, such as linkers, are more readily repaired but are more prone to lesions. Similar to nucleosomes, compact chromatin forms a physical barrier to repair proteins as well as mutagens. Consequently, the presence of divergent rates of lesion formation and DNA repair predicts evolutionary rate heterogeneity corresponding to various chromatin states.

5^mC exhibits different mutation and selection properties from other nucleotides and is one of the major causes of rate heterogeneity among sites. Previous analyses have firmly established an increased mutation rate at 5^mC because (i) it exhibits a much higher spontaneous deamination rate than normal cytosine [18]; (ii)

mismatch repair cannot distinguish between the methylated and non-methylated strands [40], which may lead to repair of T-G mismatches from 5^mC deamination on the opposite strand to produce a G→A mutation [30]; (iii) unlike U-G pair formation resulting from normal cytosine deamination that stalls DNA polymerases, T-G mismatch can be normally replicated if it persists until DNA replication. On the other hand, some 5^mC nucleotides are functionally important in gene regulation, X-inactivation or genome stability. Natural selection will oppose mutation pressure on these functional 5^mC nucleotides. Since 5^mC is confined to CpG sites in vertebrates, the joint effect of mutability and functional significance of 5^mC predicts substitution rate heterogeneity between methyl-CpG sites and other nucleotides as well as among methyl-CpG sites.

Methyl-CpG is also a strong candidate for regional differences in λ. Because deamination of 5^mC produces transition mutations, λ for methyl-CpG is expected to be higher than that for other dinucleotides. Given the association between DNA methylation and condensed chromatin structure [41, 42, 43], the methylation level has been found to be higher in tightly packed chromatin than in decondensed chromatin [44]. Thus, substitution rate heterogeneity arising from different chromatin states is derived from substitution type heterogeneity generated by methyl-CpG sites.

The effect of chromatin structure and CpG methylation on mutation rate heterogeneity will be further discussed in Chapters 3-5.

1.3.3 Other biological factors

Many of the biological factors mentioned above coexist with epigenetic factors and have impacts on evolutionary rate heterogeneity. However, their influences are distinctive and can be distinguished from heterogeneity caused by epigenetic factors.

Replication errors and male-biased mutations cause large scale evolutionary rate heterogeneity. The frequency of replication errors is affected by dNTP concentrations which vary during the cell cycle [45, 46]. Consequently, the mutation rate between each replication unit is likely to differ. Such units usually comprises 12-100 adjacent replication origins and are located 50-300 kb apart [47]. Male-biased mutation means that the mutation rate is higher in males than in females, putatively due to the larger number of germline cell divisions in spermatogenesis than in oogenesis [1]. CpG methylation [48] could also contribute to this phenomenon because DNA methylation levels are higher in male than in female germline tissues [49]. Nevertheless, male-biased mutation is considered to be the major cause of rate variation among chromosomes. Therefore, these two factors induce large scale evolutionary rate heterogeneity and should not be confounded with localised rate variations caused by epigenetic factors.

TCR specifically repairs the transcribed strand and putatively leads to substitution rate heterogeneity between genic and intergenic regions. Within genic regions, substitution rate heterogeneity caused by epigenetic factors should still be present as the non-transcribed strand is not repaired by TCR. However, since the substitution rate is expected to be largely homogeneous on the transcribed strand, the overall effect of epigenetic factors may be weakened when compared with that in intergenic regions.

1.4 Evolutionary rate measurement

The evolutionary rate, which is the rate at which genetic differences accumulate between species, is measured by the substitution rate. The substitution rate is defined as the expected number of point mutations becoming fixed in the species per generation. According to the definition, the substitution rate can be expressed by the equation [50]:

$$\theta = \mu_0 p_0 \tag{1.1}$$

where θ, μ_0, and p_0 represent the number of substitutions per generation, number of mutations per generation and the probability of fixation respectively. In a diploid population of size N, there are $2N$ gametes, so $\mu_0 = 2N\mu$ where μ is the mutation rate per generation. Under the theory of genetic drift, the probability of fixation, p_0, equals $1/2N$. Therefore,

$$\theta = \mu_0 p_0 = 2N\mu \frac{1}{2N} = \mu \tag{1.2}$$

which states that the substitution rate is equal to the mutation rate if mutations are selectively neutral. From this, it is clear that any processes that affect mutagenesis will influence the substitution or evolutionary rates.

The substitution rate can be estimated through analyses of aligned homologous sequences with a phylogenetic tree relating the sequences. Several methods have been developed to measure substitution rates, among which maximum likelihood methods are the most commonly used and are described in detail below.

1.5 A maximum likelihood approach to measure the substitution rate

Maximum likelihood estimation (MLE) is an approach that fits a probabilistic model to data. In the phylogeny-based approach, the model includes a phylogeny and parameters that represent the process of sequence divergence. In this approach, substitution models are applied to describe the process of evolutionary change and compute the likelihood of given current biological sequences.

1.5.1 Substitution models

For all the cases considered here, evolutionary processes are described under a continuous time finite Markov chain. This representation of sequence divergence is central to most molecular evolutionary analyses, including phylogenetic reconstruction [51], sequence alignment [52] and identifying the influence of natural selection [53, 54, 55, 56].

The substitution rate is derived from time T and the rate of base exchanges. We use a matrix $\mathbf{P}(T)$, whose entries p_{ij} are the probability of replacement of base i by j after time T, to represent the substitution process. Under the Markov process, we have

$$\mathbf{P}(T + dT) = \mathbf{P}(T)(\mathbf{I} + \mathbf{Q}dT) \tag{1.3}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{Q}$ is the instantaneous substitution rate matrix. $\mathbf{Q}$ has entries q_{ij} which represent the rate of replacement of i by j. This equation is solved by

$$\mathbf{P}(T) = e^{(T\mathbf{Q})} = \mathbf{I} + T\mathbf{Q} + \frac{T\mathbf{Q}}{2!} + \frac{T\mathbf{Q}}{3!} + \cdots \tag{1.4}$$

Since T and $\mathbf{Q}$ are confounded, time is expressed as the expected number of substitutions per site by scaling $\mathbf{Q}$ such that $\sum \pi_i q_{ij} = 1$, where $i \neq j$ and π_i is the equilibrium frequency of base i.

The Markov process is assumed to be stationary, time-reversible and time-homogeneous. Stationarity means that the frequencies of nucleotides are at equilibrium and do not change over time. The condition of time-reversibility means that there is no difference between creation and destruction of a base, so the rate A→G is the same as the rate G→A. This is achieved by satisfying the condition $\pi_i q_{ij} = \pi_j q_{ji}$. Consequently, **Q** is symmetric across the leading diagonal. A time-homogeneous process means that evolution for the period of time T is correctly described by a single **Q**.

Different instantaneous **Q** matrices provide different weights to exchanges of each type of base replacement. For nucleotide substitution models, **Q** is a 4x4 matrix. For instance, with the introduction of maximum likelihood estimation [51], Felsenstein (1981) presented a relatively simple model (denoted as F81) that incorporated the equilibrium frequency of nucleotides and an equal rate of substitution for all nucleotide changes (Figure 1.1). Hasegawa, Kishino and Yano (1985, HKY model) [57] extended the F81 model by considering different rates for transitions and transversions. This was represented by a single parameter λ (Figure 1.1). The extensions of rate matrices to dinucleotide and codon substitution models will be described in Chapters Three and Five.

F81

$$\begin{bmatrix} & A & G & C & T \\ A & - & \pi_G & \pi_C & \pi_T \\ G & \pi_A & - & \pi_C & \pi_T \\ C & \pi_A & \pi_G & - & \pi_T \\ T & \pi_A & \pi_G & \pi_C & - \end{bmatrix}$$

HKY

$$\begin{bmatrix} & A & G & C & T \\ A & - & \lambda\pi_G & \pi_C & \pi_T \\ G & \lambda\pi_A & - & \pi_C & \pi_T \\ C & \pi_A & \pi_G & - & \lambda\pi_T \\ T & \pi_A & \pi_G & \lambda\pi_C & - \end{bmatrix}$$

Figure 1.1: **Q matrices for the F81 and HKY models**. The diagonal elements are specified by the constraints that the rows sum to 0.

1.5.2 Maximum likelihood estimation

The likelihood is proportional to the conditional probability of the observed data given the evolutionary model. Here, the data are aligned homologous sequences. For the continuous-time Markov process models introduced above, the evolutionary model includes the relationship between sequences described as a phylogenetic tree with branch lengths, exchangeability parameters (such as λ) in $\mathbf{Q}$ and base state probabilities (nucleotide, dinucleotide or codon frequencies).

I first consider the likelihood of a single alignment column under a nucleotide substitution model. Suppose there are four species (a-d) with a known unrooted tree topology (Figure 1.2). The observed nucleotides at the kth alignment column are TCCA, and the bases at the ancestral nodes are represented by X_e and X_f respectively. The branch lengths are denoted by t_m (m = 1,2,...,5)(Figure 1.2). The four assumptions, namely time-homogeneity, rate-homogeneity, reversibility and stationarity, are applied to the whole tree topology. Thus, a single $\mathbf{Q}$ is used throughout the time dimension; and the observed nucleotide frequencies are employed at the internal nodes. Therefore, the likelihood of the kth site is written as:

$$l_k = \pi_{X_e} P_{X_e A}(t_1) \pi_{X_e} P_{X_e T}(t_2) \pi_{X_e} P_{X_e X_f}(t_5) \pi_{X_f} P_{X_f C}(t_3) \pi_{X_f} P_{X_f C}(t_4) \qquad (1.5)$$

where π_x is the nucleotide equilibrium frequency, and $p_{ij}(t)$ is computed from q_{ij} and t as described above. As X_e and X_f are unknown, the likelihood will sum over all possible nucleotide combinations for the common ancestors as:

$$L_k = \sum_{X_e=1}^{4} \sum_{X_f=1}^{4} l_k \qquad (1.6)$$

To calculate the full likelihood of the alignment, each aligned position is assumed to evolve in an independent and identically distributed manner. Thus, with the

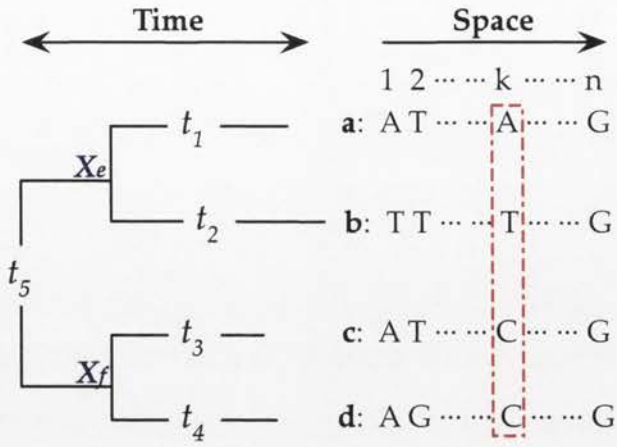


Figure 1.2: **Unrooted phylogenetic tree for four species a-d.** The time dimension represents the evolutionary time and the space dimension represents the position in the alignment. X_e and X_f are ancestral sequences of the kth column at the internal nodes. t_1-t_5 are branch lengths that represent substitution rates. The time dimension represents the evolutionary history and the space dimension represents aligned positions along the sequence.

space-homogeneity assumption, the same **Q** is applied to all alignment columns and with the independence assumption, the likelihood of an alignment is the product of the likelihood of each alignment position. Conventionally, we use the log likelihood which is given by:

$$lnL = \sum_{k=1}^{n} lnL_k \tag{1.7}$$

Finally, numerical optimisation techniques are applied to identify the vector of parameter values that maximise the likelihood, which is taken as the maximum likelihood estimated parameter values. Note that if not otherwise specified in the subsequent chapters, a gap is treated as an ambiguity character (N) to allow likelihood calculations without modeling a gap as an extra state.

1.5.3 Hypothesis testing

Different hypotheses about the evolution of the sequences can be easily formed and tested with different substitution models. As described above, prior assumptions are made to construct the likelihood function for the data. When they are violated by natural biological properties that affect substitutions, the confidence in the inferences of the substitution model is reduced. Thus, a more general substitution model, achieved by relaxing the violated constraints, will provide a better estimate with increased likelihood. Statistical tests are then performed on the difference in the likelihoods of the two substitution models to find whether the general model provides a significantly better explanation of the data. In all the cases examined in the following chapters, nested hypotheses are used. Nested hypotheses mean that the null (constrained) hypothesis H_0 with a sufficiently large amount of data is a special case of the alternative (general) hypothesis H_1. In such a situation, the likelihood ratio statistic $LR = 2ln(L_1/L_0) = 2(lnL_1 - lnL_0)$ approximately follows a χ_i^2 distribution where i is the difference of numbers of free parameters between H_1 and H_0. Therefore, a p value from a likelihood ratio test (LRT) is used to determine whether to reject or accept the null hypothesis.

Spatial substitution rate heterogeneity can be assessed by a LRT. As described above, the same evolutionary process is generally assumed at each aligned position in the sequences. Since substitution rate heterogeneity has been identified at various genomic scales, this assumption may not be satisfied for observed data. An intuitive solution is to allow a different substitution rate (t) at different segments of the alignment. For instance, the first half of the alignment evolves at rate t_1 and the second half of the alignment evolves at rate t_2. The null (H_0) hypothesis is that there is no substitution rate heterogeneity in the alignment, so there is one t in the model; and the alternative (H_1) hypothesis is that there is rate heterogeneity t_1 and t_2. H_1 becomes H_0 when constraint $t_1 = t_2$ is specified. Therefore, a LRT based on a χ_i^2 distribution is used. This approach is applied in

Chapter Three to test substitution rate and substitution type heterogeneity. Additionally, spatial heterogeneity can be formally tested by more advanced models that combine a phylogenetic model with a hidden Markov model. This will be described in Chapter Four.

Mutation properties, as well as selection properties, of specific base (nucleotide, dinucleotide and trinucleotide) exchanges can be modeled by proper parameters in substitution models and evaluated by a LRT. For instance, the F81 model assumes equal rates for all nucleotide substitutions. This is frequently violated with observations that transitions occur more often than transversions [23, 17], so the HKY model incorporates parameter λ to represent this difference. When λ equals to 1, the HKY model becomes the F81 model meaning that the transition rate equals the transversion rate. Therefore, whether the transition rate differs from the transversion rate in the sequences can be assessed by LRTs using the null hypothesis as the F81 model and the alternative hypothesis as the HKY model. Analogous tests are extensively applied in Chapter Five to test methyl-CpG mediated mutation and selection properties.

Chapter 2

Programmatic Sampling of Genomic Data

Abstract

Sampling of genomic data, including sequences, annotations, and alignments is a prerequisite for comparative genomic analyses. A software library to facilitate programmatic access to the vast quantity of data in the Ensembl MySQL databases was developed. The resulting ensembl module provides end users a simple but effective interface to approach complex data models. Programs that use the ensembl module rather than flat files to obtain data are more efficient and enable clearer expression of the data sampling procedure. This module was incorporated in PyCogent release 1.3.

2.1 Motivation

To conduct comparative genomic analyses, I needed to sample a large amount of genomic data. These data included genomic sequences, multiple species alignments, and annotations such as genes, transcripts, repeats, and SNPs. A common

approach to obtain data is to download text files from various database FTP sites. This method is straightforward, but writing parsers can be time-consuming and is not flexible enough to meet the diverse requirements of numerous users. Instead, I wanted to query and retrieve data using a "natural" biological expression that clearly states how the data are being sampled. For example, to obtain an *IL13* gene alignment for primates, the code

```
>>> human_genome = Genome("human")
>>> Il13 = human_genome.getGenesMatching("IL13")
>>> compara = Compara(["human", "macaque", "chimpanzee"])
>>> syntenic_region = compara.getSyntenicRegion(
...                   ref_species="human",
...                   location=Il13.Location)
>>> aln = syntenic_region.getAlignment()
```

is succinct and self-explanatory. This can be achieved by computer programs that interact with source databases, where high level objects like "Genome" and "Compara" are exposed to users to perform various sampling functions.

Genomic data are available from public databases, among which Ensembl and UCSC Genome Browser Database (hereafter, UCSC) are most commonly used for vertebrates. They store and display genomic sequences and features from other resources, as well as their own in-house analyses. Their data are mainly organised by species. Under each species, genomic sequences and features such as genes, CpG islands, repeats and SNPs are available. Ensembl and UCSC both use the MySQL database engine. In UCSC, each species has its own database including genome information and comparative data. In Ensembl, each species has multiple databases for different types of features. For example, gene-related data are stored in a core database, and SNPs are in a variation database. Ensembl stores comparative data from all species in a single compara database. Both Ensembl and UCSC provide online data-mining tools to facilitate customised data

download. For instance, with the UCSC table browser, a user is able to specify certain conditions for the data set and output it as a text file. A similar tool in Ensembl is called BioMart. However, at the start of my project, neither site provided multiple alignments and obtaining such alignments required significant post-processing of text files to join information from different files.

Although providing largely the same underlying genomic and annotation data, Ensembl and UCSC differ significantly in the nature of their genomic sequence alignments resulting in my choice of Ensembl. Because genomic sequence alignment is such a computationally challenging problem, it was more feasible to take advantage of publicly available results from groups already working on this problem. Thus, the availability of ready-to-use alignments was a critical requirement. In UCSC, alignments are provided in MAF format where genomic coordinates and aligned sequences are in a single line. It breaks whole genome (or chromosome) alignments into small blocks that are usually hundreds to thousands of base pairs long. While genomic coordinates for reference species are continuous among aligned blocks, they are not for other species. For example, to obtain a long alignment for human-mouse-rat from the human database, sequentially connecting records only works for human sequences, while parts of mouse and rat sequences are missing between blocks. Constructing a correct alignment involving full sequences from human, mouse and rat can therefore be time-consuming. By contrast, Ensembl first identifies syntenic regions among species followed by multiple genome alignment. Ensembl alignments are in large blocks with full aligned sequences available for all species. Second, the relationships of proteins between and within species are clearer in Ensembl than UCSC. For instance, Ensembl specifies ortholog proteins as having one-to-one, one-to-many, and many-to-many relationships, while such information is ambiguous in UCSC. Third, Ensembl regularly updates its databases with a release number, which makes it easy to track changes. Such information in UCSC is unclear, for instance, there is only one human database version hg18 from March 2006 to February 2009. For these

reasons, I selected the Ensembl databases as the primary source for my research data and developed a library for sampling data from them.

An application that directly queries and retrieves data from Ensembl MySQL databases was developed in the Python programming language. MySQL is a relational database based on different entities (e.g. tables) and the relations among them. Through Structured Query Language (SQL), MySQL provides much better performance than flat file-based storage systems. Python is able to access MySQL through the Python library MySQLdb. Since the subsequent evolutionary analysis code in this thesis is in Python [58], using the same language for database querying and evolutionary analyses facilitates the development of an efficient, consistent and reliable research workflow. At the time of starting my project, there was no programmatic access to the Ensembl database in Python.

The Ensembl Python module was developed and integrated with PyCogent. The design of this module was determined by writing use-cases, examples of what were deemed queries necessary for completion of my project. These use-cases were written as Python doctest documents and the final version of this design document is now available as part of the PyCogent distribution. As end-users of a database designed by others, the execution of queries necessarily follows the connection of database entities, forming completed information by joining data distributed over multiple tables. The resulting objects express genomic data in what is viewed as their natural relations. Since all the genomic data were from the same source data, details such as genomic coordinates and strands are consistent. The code objects resulting from queries were standard PyCogent objects whose attributes were easy to understand and manipulate, greatly facilitating subsequent sampling and analysis procedures. Note that by convention, a different font is used for the names of Ensembl entities and PyCogent objects below.

2.2 Features

2.2.1 Design

The design objectives were to: present the genomic content in a biological intuitive and succinct way; be applicable to remote or local installations of Ensembl; be computationally efficient; and to be readily maintainable. Sampling procedures were simplified such that a complete data set could be produced with relatively few lines of code. Such a design, accompanied by adherence to good coding principles, ensures the intent of a script is legible even to a non-programmer. This will be illustrated in detail below.

2.2.2 Ensembl database schema

A prerequisite for retrieving data from Ensembl databases is to understand the relations between different entities and how to locate relevant information. In a relational database, information is separated, stored in different tables, and connected to each other through a reference index. For example, there are several tables associated with sequence information in the core database, including (i) the `coord_system` table that records the level of genomic coordinate, such as supercontig or chromosome, (ii) the `seq_region` table that records the name of a sequence region for each level (e.g. 1, 2,..., 22, X, Y under the chromosome level), (iii) the `assembly` table that converts coordinates between different coordinate systems (e.g. between contig and supercontig). The first two tables are connected by an integer index `coord_system_id`. In this way, storage space is saved, because instead of iterating `coord_system` table records in the `seq_region` table, an integer is used to reference the related records. Another advantage is that it is more convenient when modifying the stored data. For instance, if we update supercontig to ultracontig, only one record is changed in the `coord_system` table, while the `seq_region` table remains the same. Consequently, querying involves

joining relevant tables to obtain complete information.

A schema view of the Ensembl variation database is shown in Figure 2.1. Compared with the Ensembl core database, the variation database is much less complex. In the figure, boxes represent entities, which can be a single table or a group of related tables; and lines represent relations between tables. Each table has a primary key that is a unique identifier for each record in the table. The primary key can be referenced from other tables. Tables may also have a foreign key which matches the primary key column of another table to connect to other entities. Using the same example above, `coord_system_id` is the primary key in the `coord_system` table, but is the foreign key in the `seq_region` table. Altogether, there are hundreds of tables in different databases, and, as Figure 2.1 demonstrates, their relations are complex.

To make entity relations clearer, I identified the tables from a database that contained data pertinent to a specific biological concept and built up a smaller scale schema view. Figure 2.2 is one example that shows the gene-related tables in the core database. Briefly, a gene has one or more transcripts, each of which has one or more exons. The corresponding data are stored in the `gene`, `transcript` and `exon` tables respectively. For protein-coding genes, the `translation` table records translation start and end exons corresponding to a transcript, so it connects to the `transcript` table through `transcript_id`. The `exon_transcript` table identifies an exon belonging to a transcript, whereas associated stable ids and symbols are all stored in separate tables. Tables with `seq_region` columns are associated with sequence information. This procedure was used to identify relations in Ensembl databases and locate the required information.

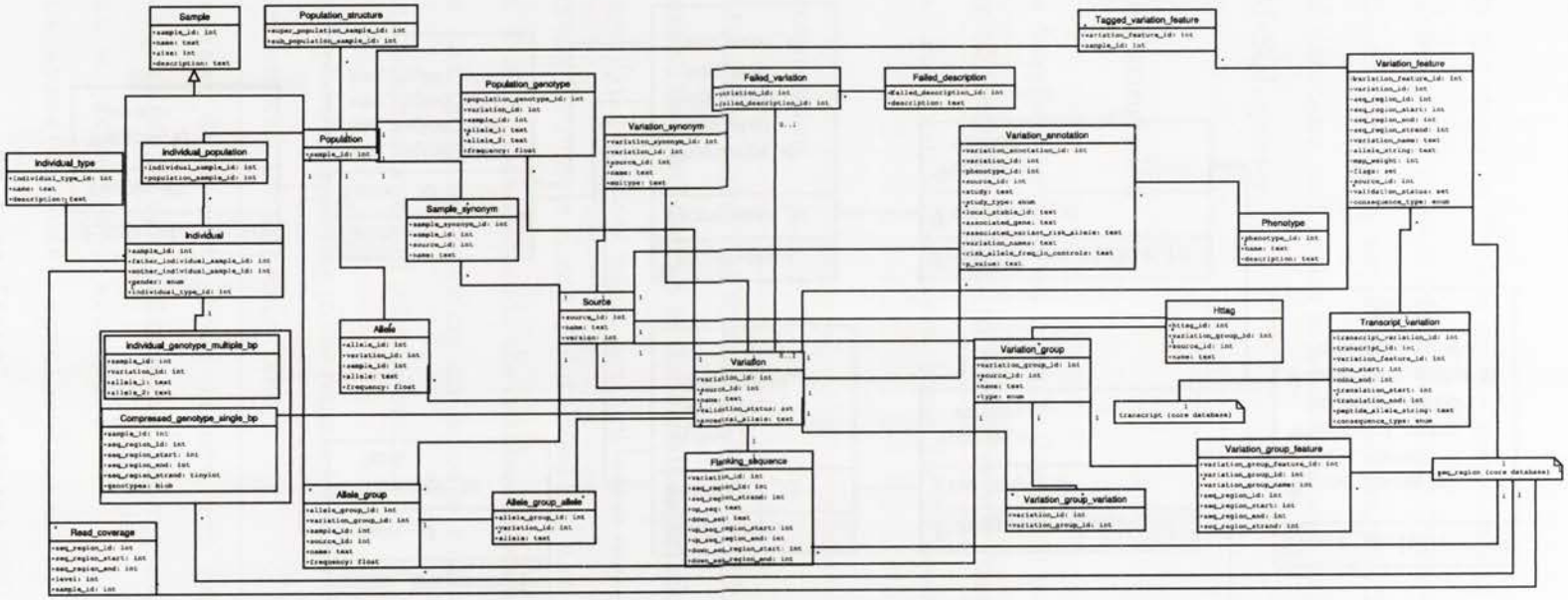


Figure 2.1: **Ensembl variation database schema, release 55.** It was downloaded from `http://www.ensembl.org/info/docs/api/variation/variation-database-schema.pdf`. Boxes represent entities, which can be a single table or a group of related tables. Lines represent relations between tables.

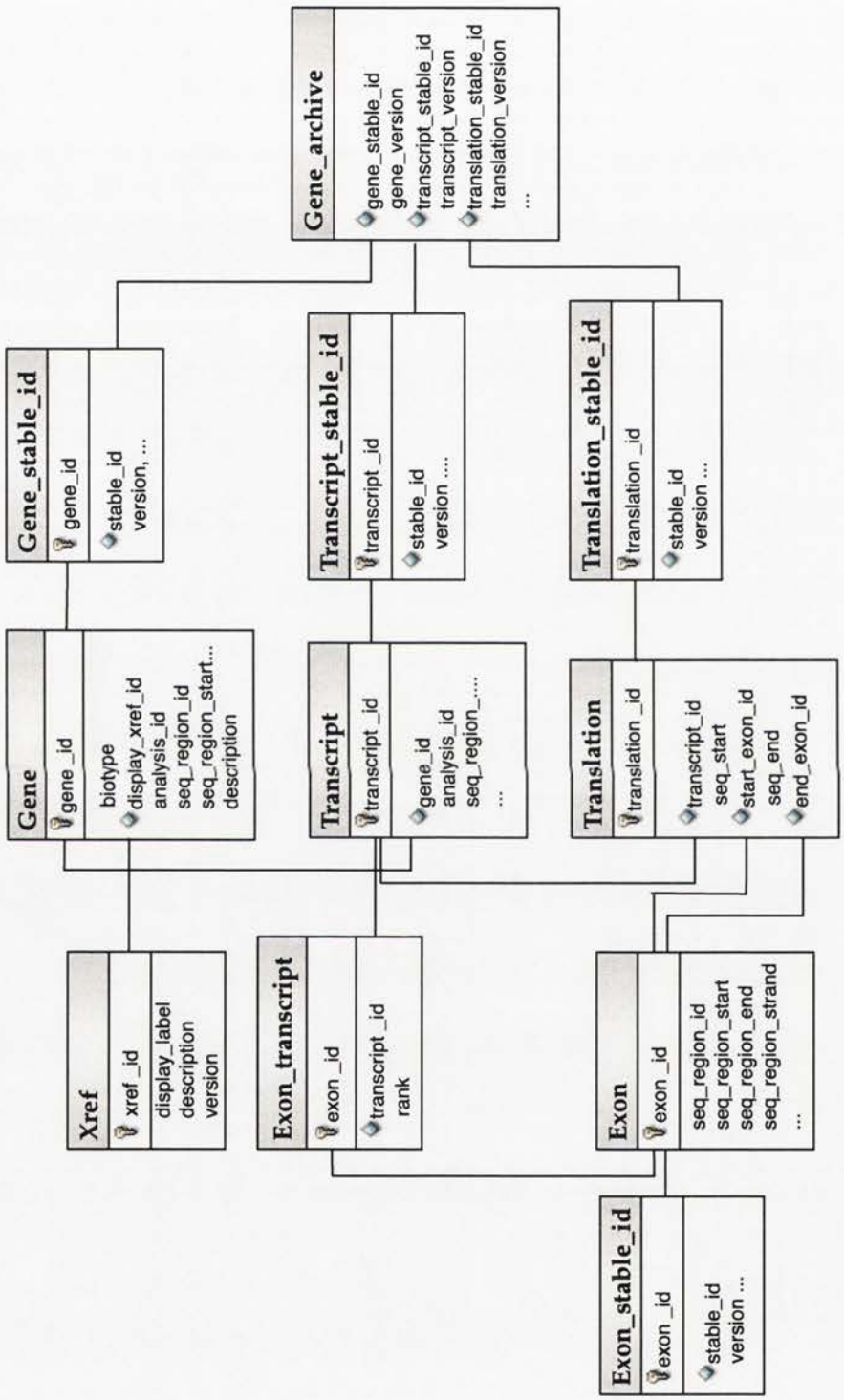


Figure 2.2: **Ensembl core database schema for gene-related tables** Each box represents one table in the Ensembl core database. The first row of the box is the table name; the second row with a key mark is the primary key in the table; and the third row lists some column names with foreign key(s) identified by a cyan diamond mark.

2.2.3 Implementation

Having made the choice of Python as the programming language, we further selected Sqlalchemy, a Python module, to handle database-related issues. It is a well-established third-party application that controls the connection and disconnection of databases and tables in a sophisticated way, and allows dynamic building of MySQL queries.

2.2.4 Key modules

The modules described here are key modules developed to perform various functions. Many of them are hidden to end users who usually deal with higher level objects. For sophisticated users, these are all accessible for their special needs.

Database connection

Data retrieval starts by connecting to a database, located locally or remotely, with a valid account. The `HostAccount` in the host module allows a user to create an account by specifying the host name, account name and password. The default Ensembl account for remote Ensembl databases can also be imported from the host module. The `Database` object in the database module provides a connection to databases and tables by providing an account with optional arguments of species name, database type and release version.

For example, explicitly connecting to the remote UK Ensembl human core database can be carried out with the following code:

```
>>> from cogent.db.ensembl.host import get_ensembl_account
>>> from cogent.db.ensembl.database import Database
>>> account = get_ensembl_account()
>>> db = Database(account, species="human",
...               db_type="core", release=52)
```

However, this is made more convenient for users by setting the UK servers as the default account, and providing a wrapper `Genome` class (described below).

Name handling

Ensembl database names for each species start with their Latin names followed by the database type, release version and database build. Because users are likely to use a species common name, the `Species` object in the species module offers a map between the Latin and common names for a species. For instance,

```
>>> from cogent.db.ensembl.species import Species
>>> name = Species.getSpeciesName("human")
>>> print name
Homo sapiens
```

Default species names that are accepted by a `Genome` object (below) can be viewed by directly printing the `Species` object.

```
>>>print Species
```

Common Name	Species Name	Ensembl Db Prefix
...	...	...
Cat	Felis catus	felis_catus
Chicken	Gallus gallus	gallus_gallus
Chimp	Pan troglodytes	pan_troglodytes
Cow	Bos taurus	bos_taurus
...	...	...

If a species name is not in this table but is valid in Ensembl, it can be added through an `amendSpecies` method. A `Genome` instance can then be constructed using the user-defined species common name as usual (see below).

The Genome object

A `Genome` is an important object designed for interrogating genomic data. With the exception of comparative data, it is capable of performing queries on all databases related to a species. A `Genome` object is instantiated using a species name, with the Ensembl release number and an optional account argument. If the optional argument is not specified, the UK Ensembl account will be used. For example, the following lines create a `Genome` instance for the human genome from release 52:

```
>>> from cogent.db.ensembl import Genome
>>> human = Genome("human", release=52)
```

From `Genome`, users can search for a gene with a symbol, or get genes of a certain type as:

```
>>> brca1 = list(human.getGenesMatching(Symbol="BRCA1"))[0]
>>> genes = human.getGenesMatching(BioType="protein_coding")
```

Optional arguments for the `getGenesMatching` function are Symbol (typically the HUGO Gene Nomenclature Committee (HGNC) name), StableId which is the Ensembl gene stable identifier, BioType which is the type of gene product and Description that comprises the HGNC long gene name, alternative titles and / or source.

To get a genomic region, the `getRegion` method will return a `Region` object which is described below.

```
>>> brca1_region = human.getRegion(region=brca1)
>>> print brca1_region
generic_region(Species="Homo sapiens"; CoordName="17";
    Start=38449839; End=38530994; length=81155; Strand="-")
```

Optional arguments of the `getRegion` function are CoordName, Start, End etc. that are used to define a genomic coordinate.

To get a list of the annotated features within a region, the `getFeature` method can be used as follows:

```
>>> features = list(human.getFeature(region=brca1_region,
...          feature_types=["gene", "cpg"])
>>> for feature in features:
>>> ...    print feature, "\n"

gene(Species=Homo sapiens; BioType="protein_coding";
     Description="Breast cancer  type...";
     Location=Coordinate(
         "Human":"chro...":"17":38449839-38530994:-1);
     StableId="ENSG00000012048"; Status="KNOWN";
     Symbol="BRCA1")

CpGisland(CoordName="17"; Start=38526029; End=38526480;
          length=451; Strand="-", Score=108.0)
```

Valid feature types are Est, gene, variation, repeat and CpG (CpG island). For a gene feature query, only `Gene` objects are returned, from which features like transcripts and exons can be obtained.

Other useful methods include `getEstMatching`, `makeLocation`, `getVariations` and so on. The full capabilities and instructions for use of any of the `cogent.db.ensembl` objects can be found using the Python dir() and help() functions.

Region features

A `Region` is a bridging object that represents portions of a genome. Its key capabilities are to get sequences, features, and annotated sequences. Inherited

from the Region object, GenericRegion is used for simple features like CpGisland and repeat; while StableRegion which inherits from the Region object is used for features with an Ensembl Stable ID, such as gene, transcript, exons, SNPs and proteins. As illustrated above, many Genome functions return Region objects that can be used to extract more information. Continuing from the above example,

```
>>> brca1 = list(human.getGenesMacthing(Symbol="BRCA1"))[0]
```

brca1 is a Gene object inheriting from StableRegion that is inherited from the Region object. Therefore, the *BRCA*1 gene stable Id is:

```
>>> print brca1.StableId
ENSG00000012048
```

From a Gene object, a Transcript object can be obtained from the attributes of Transcripts or CanonicalTranscript.

From the Transcript object, Exons are ready to use as follows:

```
>>> brca1_transcript1 = brca.Transcripts[0]
>>> exons = brca1_transcript1.Exons
```

Other features like CDS, 5'UTR, etc. can be obtained in a similar way.

A Compara database query

A Compara is the primary object designed to retrieve comparative genome data. It has the capacities to connect to the Ensembl compara database, connect to the genome databases of each individual species, and get orthologs and syntenic regions. Similar to Genome, it is initiated with a valid set of species names, while the arguments of account and Release are optional.

```
>>> from cogent.db.ensembl import Compara
>>> compara = Compara(["human", "mouse", "rat"], Release=52)
```

A `Genome` object is accessible from `Compara` as follows:

```
>>> human_genome = compara.Human
```

There are two major functions currently implemented, one of which is to get orthologs and the other is to get syntenic regions. For example, to get human *BRCA1* orthologs in mouse and rat is simply as follows:

```
>>> orthologs = compara.getRelatedGenes(gene_region=brca1,
    ...                 Relationship="ortholog_one2one")
```

which returns a collection of `Gene` objects for each species queried, if it is available. Valid relationships are defined by Ensembl as ortholog_one2one, ortholog_one2many, between_species_paralog and so on. Getting syntenic regions from Compara is also straightforward:

```
>>> syntenics = compara.getSyntenicRegions(region=brca1,
    ...          align_method="ORTHEUS",
    ...          align_clade="9 eutherian mammals")
```

A valid alignment method and clade can be found from the Ensembl website. The `SyntenicRegion` object allows the retrieval of sequence alignments. The use of these two functions will be further illustrated in the following examples.

Unittest

Each module was thoroughly tested to validate the accuracy of extracted information. Not only functions, but also major concepts, such as reverse complements, strand information, and coordinates, were carefully tested. Sequences and implemented features were checked against data downloaded from the Ensembl website to verify correct resolution of genomic coordinates and handling of annotations. Whenever a new database version is released, the unittests are run to ensure that each module continues to work properly - evidenced by all tests passing.

2.3 Examples

The capabilities of the PyCogent ensembl module are substantial, but here I just illustrate some cases directly relevant to the genomic sampling carried out in the following chapters. For more information, the online documentation (`http://pycogent.sourceforge.net/examples/query_ensembl.html`) provides more details of the design features and the ensembl test module displays more usage examples.

2.3.1 Example 1: getting a sequence with annotations for a genomic coordinate

.

For a given genomic coordinate, a `Region` object can be generated by the `Genome` `getRegion` method. From the returned region, the `getAnnotatedSeq` function returns a sequence with the required annotations. Thus, the following code suffices:

```
>>> from cogent.db.ensembl import Genome
>>> human = Genome("human", Release=54)
>>> start = 29656651
>>> region = human.getRegion(CoordName=20, Start=start,
    ...                      End=start+1999, Strand=1,
    ...                      ensembl_coord=True)
>>> seq = region.getAnnotatedSeq(
    ...              feature_types=["gene", "cpg", "repeat"])
>>> for annot in seq.annotations:
>>> ...    print annot

gene "ID1" at [102:1324]/2000
```

```
transcript "ENST00000376105" at [102:1324]/2000
exon "ENSE00001469386-1" at [102:1324]/2000
CDS "ENST00000376105" at [201:651]/2000
5'UTR "ENST00000376105" at [102:201]/2000
3'UTR "ENST00000376105" at [651:1324]/2000
transcript "ENST00000376112" at [102:1324]/2000
exon "ENSE00001469419-1" at [102:627]/2000
exon "ENSE00001469404-2" at [866:1324]/2000
CDS "ENST00000376112" at [201:627, 866:908]/2000
5'UTR "ENST00000376112" at [102:201]/2000
3'UTR "ENST00000376112" at [908:1324]/2000
CpGisland at [1036:0, -135-]/2000
repeat "trf" at [358:334]/2000
repeat "MER5A" at [1616:1579]/2000
repeat "MIR3" at [1771:1683]/2000
```

The purpose of the ensembl_coord argument is to handle the difference between the Ensembl index, which starts from 1, and the Python index, which starts from 0. By default, gene features include gene-related features, such as transcripts, exons, UTRs, and so on. Sequence annotations record feature types, labels and positions including start, end and missing parts relative to the current sequence.

2.3.2 Example 2: retrieving all human protein-coding genes and their coding sequences from canonical transcripts

Protein-coding genes are one particular gene type, so specifying 'BioType = protein_coding' in the `getGenesMatching` function will obtain all protein-coding genes. Canonical Transcript is an attribute of the `Gene` object, and CDS is an attribute of `Transcript`. Hence, code is as simple as follows:

```
>>> from cogent.db.ensembl import Genome
>>> human = Genome("human", Release=52)
>>> genes = human.getGenesMatching(BioType="protein_coding")
>>> for gene in genes:
>>> ...    cano_transcript = gene.CanonicalTranscript
>>> ...    cds_seq = cano_transcript.Cds
>>> ...    print "Gene:%s; Transcript:%s; CDS=%s..."%
...       (gene.Symbol, cano_transcript.StableId,
...        cds_seq[:10])

Gene:RBM15; Trancript:ENST00000369784; CDS=ATGAGGACTG...
Gene:MAP3K6; Trancript:ENST00000357582; CDS=ATGGCGGGGC...
Gene:OR4F5; Trancript:ENST00000326183; CDS=ATGGTGACTG...
Gene:OR4F29; Trancript:ENST00000327169; CDS=ATGGATGGAG...
Gene:OR4F16; Trancript:ENST00000332831; CDS=ATGGATGGAG...
... ...
```

2.3.3 Example 3: retrieving gene sequence alignments on human chromosome 1 for human, macaque and chimpanzee

This is a comparative task, so the first step is to create a `Compara` object with three primates. The reference species is human whose `Genome` object can be accessed through the 'Human' attribute of `Compara`. From the human genome, a region of chromosome 1 and its genes can be obtained. Therefore, the code is as follows:

```
>>> from cogent.db.ensembl import Compara
>>> compara=Compara(["human","chimp","macaque"], Release=52)
>>> # connect to human Genome object
```

```
>>> human = compara.Human
>>> human_chr1 = human.getRegion(CoordName=1)
>>> chr1_genes = human_chr1.getFeatures(feature_types="gene")
>>> for gene in chr1_genes:
>>> ... syn_regions = compara.getSyntenicRegions(region=gene,
    ...          align_method="ORTHEUS", align_clade="primates")
>>> ... # convert generator to list
>>> ... syn_regions = list(syn_regions)
>>> ... # if no syntenic regions in the other two species
>>> ... if not syn_regions:
>>> ...     continue
>>> ... # if have syntenic regions:
>>> ... print "Gene:%s;\nStableId:%s; Type:%s"%(gene.Symbol, \
    ...          gene.StableId, gene.BioType)
>>> ... for syn_region in syn_regions:
>>> ...     aln = syn_region.getAlignment()
>>> ...     print aln[:50], "\n"


Gene:AL627309.15;
StableId:ENSG00000197490; Type:retrotransposed
>Homo sapiens:chromosome:1:42911-44799:1
CTTATATCCATAGCTACCTGTTTCTTATTAATAATATCCTATATTTTCAT
>Macaca mulatta:chromosome:7:82191434-82194390:-1
CTTATATCCATGGCTGCCTGTTTCTTATTAATTATATCCTATATTTTCAT
>Pan troglodytes:chromosome:Un:9708159-9710058:1
CTTATATCCATAGCTACCTGTTTCTTATTAATAATATCCTATATTTTCAT


Gene:AL627309.15;
StableId: ENSG00000205292; Type:pseudogene
>Homo sapiens:chromosome:1:52877-53750:1
ATGCAGTTTTTCCTTTTTCTCTTCTTCTCTTTATTCTATGTGGGAATTAT
```

```
>Macaca mulatta:chromosome:7:82182115-82182988:-1
ATGCAGTTTTTTCTTCTTCTCTTCTTCTCTTTATTCTATGTGGGAATTAT
>Pan troglodytes:chromosome:Un:9718780-9719653:1
ATGCAGTTTTTCCTTTTTCTCTTCTTCTCTTTATTCTATGTGGGAATTAT

Gene:OR4F5;
StableId: ENSG00000177693; Type:protein_coding
>Macaca mulatta:chromosome:7:82176264-82177179:-1
ATAGTGACTGAATTCATTTTTCTGGGTCTCTCTGATTCTCAGGAACTCCA
>Homo sapiens:chromosome:1:58953-59871:1
ATGGTGACTGAATTCATTTTTCTGGGTCTCTCTGATTCTCAGGAACTCCA
>Pan troglodytes:chromosome:Un:9726136-9727053:1
ATGGTGACTGAATTCATTTTTCTGGGTCTCTCTGATTCTCAGGAACTCCA
... ...
```

2.3.4 Example 4: retrieving human and mouse ortholog protein-coding genes with a one-to-one relationship

The aim of this code example is to find ortholog genes between human and mouse, which is also a comparative task. I first create a `Compara` object with human and mouse being passed to the species set argument. Again, from the human genome, `getGenesMatching` is applied to return all human protein-coding genes, which are further used to find related genes in the mouse. Thus, the code is straightforward:

```
>>> from cogent.db.ensembl import Compara
>>> compara=Compara(["human","mouse"], Release=52)
>>> human=Compara.Human
>>> human_genes=human.getGenesMatching(
    ...                     BioType="protein_coding")
```

```
>>>
>>> for gene in human_genes:
>>> ... orthologs=compara.getReleatedGenes(gene_region=gene,
    ...                 Relationship="ortholog_one2one")
>>> ... if not orthologs:
>>> ...     continue
>>> ... print "Gene %s orthologs:"%gene.Symbol
>>> ... members=orthologs.Members
>>> ... for member in members:
>>> ...     print "StableId=%s;\nLocation="%member.StableId,\
    ...                                         member.Location
>>> ... print

Gene RBM15 orthologs:
StableId = ENSMUSG00000048109;
Location= Mus musculus:chromosome:3:107128865-107135998:-1
StableId = ENSG00000162775;
Location= Homo sapiens:chromosome:1:110678038-110690818:1

Gene MAP3K6 orthologs:
StableId = ENSMUSG00000028862;
Location= Mus musculus:chromosome:4:132796732-132808843:1
StableId = ENSG00000142733;
Location= Homo sapiens:chromosome:1:27554256-27565970:-1
... ...
```

2.4 Conclusion

The ensembl module included as part of the open source package PyCogent library version 1.3 substantially enhances the utility of PyCogent for vertebrate genomics.

Programmatic access to various types of data is efficient and consistent. Scripts follow the natural relations between genomic features that clearly express how data are sampled and the returned data are in objects that are easy to interrogate, manipulate and save. All samplings from the Ensembl databases reported in this thesis were carried out using this code.

2.5 Requirements

The PyCogent-ensembl module requires the following software: Python 2.5 or greater, MySQL 5.0 or greater, Sqlalchemy 0.4.2 or greater, and MySQL-python 1.2.2 or greater.

Chapter 3

Evidence that Chromatin Compaction Affects Substitution Rate

Abstract

Evolutionary rates are not constant across genomes and chromatin structure has been identified as one of the potential causes of this, but the relationship between them remains elusive. Experimental studies have established a potential association between chromatin compaction and substitutions through DNA repair pathways and hypermutable 5^mC. I examined this relationship using regions annotated as DNase I Hypersensitive sites (DHSs, having an open chromatin structure) and their contrasting flanking sites (Flanks, having a comparatively closed chromatin structure). I sampled putatively neutral DHS and Flank sites in intergenic and intronic regions from primates. Likelihood ratio tests were conducted to compare differences in substitution rates and types between matched DHSs and Flanks. These tests revealed significant increases in total, general transition, and CpG transition substitutions in closed chromatin compared to open chromatin in intergenic regions. For intronic regions, however, only the

increase in total substitution rate was significant. These results support an association between chromatin structure and heterogeneity in the total rate and types of substitutions, consistent with reduced efficiency of DNA repair in closed chromatin.

3.1 Motivation

Both experimental and evolutionary studies suggest that chromatin status affects the processes of mutation, but the nature of its influence on sequence divergence remains unclear. Examination of spontaneous mutation rates at different genomic locations has revealed significant variations of up to 60-fold [59]. Chromatin structure is known to influence recombination, formation of DNA lesions and DNA repair. A recent molecular evolution study [8] identified a negative correlation between chromatin accessibility and the substitution rate in intronic, intergenic and repeat sequences, but a positive relationship between chromatin accessibility and synonymous substitution rates and transition rates. The authors proposed a lower rate of DNA damage, or enhanced DNA repair, as potential causes of the former, and stronger selection in closed chromatin regions at synonymous sites as a major cause of the latter. However, these inconsistent observations could also result from the confounding effects of chromatin structure, natural selection, and potentially, GC content, on substitutions arising from examining DNA sequences at Mbp scales as targeted regions.

While these studies support an association between chromatin compaction and substitution mutations, they fail to resolve further details of the nature of this relationship. Important remaining questions include: (1) At what scale does chromatin structure affect substitution rate? Besides open and closed chromatin fibres, which may be separated by ~1Mbp, changes to chromatin state also occurs on smaller scales, like DHSs versus Flanks. The question of whether substitution rate heterogeneity exists at a localised level of hundreds to thousands of bps and

corresponds with changes in chromatin state has not been addressed. (2) Are the types of substitution affected by chromatin state? Mutations from replication errors and mutagenesis are likely to have different tendencies to cause transitions or transversions. Thus, examining the types of substitution may facilitate understanding the sources of mutations. (3) Do CpG substitutions differ in different chromatin states? Most previous studies have only examined non-CpG substitutions. Because 5^mC tends to be associated with closed chromatin, its high spontaneous transition mutation rate potentially causes CpG transition substitution heterogeneity between regions with different chromatin states. Whether this functional association alone can account for the reported association between transition substitution heterogeneity and chromatin state is unknown. (4) What are the underlying causes of rate heterogeneity, DNA lesion formation, DNA repair, or natural selection? The answer to this last question may be inferred from the observed directions of the rates and the types of substitution heterogeneity between DHSs and Flanks.

To address these questions thoroughly, knowledge regarding chromatin and DNA metabolism is required. Chromatin compaction, which occurs at three levels (for a review, see [60]), serves a structural function as well as being a regulator of cellular activities. The first level is that of nucleosomes, which comprise ~147 bp DNA wrapped around a histone core. Nucleosomes repeatedly occur along DNA to form nucleosome arrays. In the next level, several nucleosomes pack together and coil into a helical 30nm fibre called a solenoid. This is further compacted to form chromatin fibres and condensed entire chromosomes. The final level of chromosome packaging greatly reduces the volume of DNA with an ~1000-fold reduction from its original length. This compaction also restricts DNA accessibility. Hence, chromosome structure potentially affects all nuclear activities that require DNA as a template, including transcription, DNA repair, replication and recombination. To overcome the structural barriers presented by compacted chromatin and facilitate various cellular processes, the relative

accessibility of chromatin is regulated by, for example, modifications to histone tails [61] and chromatin remodeling [62, 63].

Chromatin status has been demonstrated to affect DNA repair rates with compaction greatly reducing DNA repair efficiency. The DNA repair machinery requires direct DNA-protein contact to recognise and remove lesions [64, 65]. That DNA repair was much slower in genomic than free DNA [27, 28, 66] indicates an inhibitory effect of packed DNA on repair. This is consistent with compact chromatin preventing access to the repair machinery. Such an effect was further supported by the absence of DNA repair in the extremely compact chromosomes in mature spermatozoa [67, 68]. The consequence of reduced repair efficiency is increased longevity of DNA lesions, increasing the likelihood that a lesion will be converted to a mutation during the next round of DNA replication.

On the other hand, DNA damage preferentially attacks open chromatin structures. In the same way that compact chromatin makes DNA less accessible to repair, it also makes attack by DNA damaging agents less likely. As a result, more lesions have been observed in the decondensed chromatin of DHSs than in the compact chromatin [69]. Highly condensed mature spermatocyte DNA has further been shown to be resistant to benzpyrene-induced damage [70]. As a result, heterogeneity of lesion formation alone predicts more mutations in open chromatin structures. However, the relatively greater accessibility of DHS regions to repair systems should result in a comparatively lower mutation rate than that of their flanking regions [71].

Mutation rate heterogeneity caused by chromatin compaction also suggests biases in different mutation type profiles in different chromatin structures. There are two major mutational inputs, those arising from errors during DNA replication and those arising from DNA damage. Replication errors are expected to produce homogeneous substitution type bias within large scale sequences. For instance, Topal and Fresco [23] proposed that transitions were more abundant

than transversions during DNA replication due to the natural frequencies of base tautomers. In contrast, DNA damage processes differ in their tendency towards transition or transversion bias and such effects should be localised. For example, hydrolysis and UV-damage predominantly produce transition mutations, while oxidative 8-OH-dG damage induces GC → TA transversions [72]. Thus, subject to lesion types and repair efficiency, the transition to transversion ratio (λ) may differ between different chromatin states.

Another feature of chromatin compaction is its association in mammals with hypermutable 5^mC. 5^mC exhibits an approximately 10-fold accelerated rate of C → T transition mutations [73, 20, 21, 74]. Direct evidence suggests that DNA methylation triggers chromatin condensation on promoters followed by gene silencing (reviewed in [75]). In addition, DNA in compact chromatin exhibits a significantly greater density of CpG methylation [44]. Consistent with this observation, 5^mC nucleotides suppress formation of DHSs [76]. Consequently, the association between 5^mC and chromatin state predicts differential CpG context-dependent transitions between various chromatin states.

Differential mutagenesis between open and closed chromatin states predicts heterogeneity in both the rate and type of substitutions. Given that the mutation rate equals the substitution rate for neutral DNA sites, the substitution rate and types are expected to exhibit the same tendency as the mutation rate and mutation types. Thus, we conjecture that open and closed chromatin will differ in total substitution rate, substitution types and CpG transition substitutions arising from mutagenesis.

Recent genome-wide DHS mapping makes an investigation of the association between chromatin state and variation in sequence divergence possible. DHSs are DNA regions hypersensitive to DNase I cleavage. DHSs were first detected in SV40 viral chromatin and later found in eukaryotes as an essential feature of chromatin structure (reviewed in [77]). Usually a DHS is a few hundreds of nu-

cleotides in length, but may extend for thousands of base pairs. Examination of the chromatin structure within DHSs revealed long nucleosome-free regions and/or unusual nucleosome structures [78, 79, 76]. Conventionally, Southern blotting was employed with DNase I-treated genomic DNA to detect DHS locations [80]. More recently, high-throughput methods have been developed to identify DHSs at a high resolution [81], including both microarray and high-throughput sequencing. These techniques are being utilised to provide a detailed map of DHS distribution in the human genome [82]. The traditional view of DHSs as locus control regions that incorporate transcriptional regulators and that are mostly located close to genes has been challenged by the detection of a substantial proportion of DHSs that are far away from genes. Currently, these regions have no known functional roles. The recent development of genome-wide DHS mapping has made it possible to explore in this thesis the influence of chromatin accessibility on substitutions on a localised scale.

In the remainder of this chapter, I examine the relationship between chromatin structure and sequence variation, including both total rate and types (represented by λ) of substitutions. By contrasting DHSs (representing open chromatin) and Flank regions (flanking sites representing closed chromatin), the following questions are addressed: (1) Does chromatin structure affect total substitution rate? (2) Does chromatin structure affect substitution types? (3) Does chromatin structure affect CpG substitutions? (4) Is DNA repair the major cause of differences in substitutions between different chromatin structures?

3.2 Methods

3.2.1 Data

Ensembl Release 50 was used to obtain genomic features and multiple genome sequence alignments. The genomic features selected were genes and CpG islands.

Unless specified, genes included protein-coding genes and other types of genes, e.g. snRNAs, rRNAs, and pseudogenes. ORTHEUS genomic multiple sequence alignments of primates were sampled using human genomic coordinates [83].

DHS NCBI35 coordinates, previously defined by Boyle et. al. [82], were downloaded from UCSC using the table browser and converted to NCBI36 assembly coordinates using the LiftOver tool [84]. In order to minimise the number of potential regulatory elements within a DHS, DHSs with a length between 300 and 2000 base pairs were retained. Coordinates of the remaining DHSs were used to query the Ensembl databases to find their locations relative to genes. To further reduce the influence of natural selection, only intronic and intergenic DHSs were selected. Additionally, only intergenic DHSs located more than 3000 base pairs either side of Ensembl protein-coding genes were used.

For selected DHSs, matching non-DHS control regions (Flanks) were obtained by extending the coordinates both upstream and downstream. The Flank was sampled such that its total length was equal to that of its adjacent DHS region. For intergenic regions, a Flank was sampled so the lengths of 5'-Flank and 3'-Flank were identical. A similar strategy was applied to intronic Flanks if neither the 5'- or 3'-Flanks overlapped exons. For Flanks that spanned an exon, the exon was excluded and the length of the intronic side Flank expanded to maintain equal lengths of DHSs and Flanks. Intronic DHSs whose length was greater than half of the length of the associated intron were excluded.

Multiple sequence alignments from human, chimpanzee and macaque were sampled using the DHS + Flank coordinates of human sequences. Alignments with more than 10% gaps or N's were eliminated from the sampled data. In addition, to avoid violation of the phylogenetic models caused by extreme compositional heterogeneity along alignments, alignments with annotated CpG island sequences were excluded. Following this, 6,705 intergenic and 7,150 intronic alignments remained. To eliminate the influence of non-point mutation processes (such as

insertions and deletions) and ambiguous DNA sequences on the estimation of substitution rate, alignment columns containing a non-nucleotide character were discarded. The resulting alignments were used for analyses.

3.2.2 Software

All evolutionary modeling was done using PyCogent version 1.3.0.dev [58] and all scripts were written in the Python programming language.

3.2.3 Statistics

Likelihood ratio tests (LRTs) were conducted to compare evolutionary rate parameters between DHSs and Flanks. As described in Chapter One, the LRT is a statistical method for measuring support for two competing hypotheses. In the case of parameter comparisons, the alternative hypothesis allowed the parameters being compared to differ between DHSs and Flanks, while the null hypothesis specified that these parameters were equal between DHSs and Flanks. A significance level of 5% was used for rejecting the null hypothesis.

Substitution rate and transition/transversion ratio (λ) differences between DHSs and Flanks were assessed by LRTs under the HKY model. The substitution rate was measured as the branch length (k), defined as the expected number of substitutions per site. For an unrooted phylogenetic tree relating human, macaque and chimpanzee, there are three independent branch lengths. The total substitution rate was defined as the sum of these branch lengths ($K = \sum k$). To compare substitution rates, the alternative model was that the branch lengths differed between DHSs and Flanks ($K_{\text{DHS}} \neq K_{\text{Flank}}$). Then the likelihood function was constructed under the two models and maximized using the PyCogent built-in numerical optimizers at the default settings. With an additional set of branch lengths in the alternative model, the differences in degrees of freedom (df) are 3.

If $p \leq 0.05$, the substitution rate was considered nominally significantly different between the two compared regions. For nominally significant cases, a $K_{\text{DHS}} < K_{\text{Flank}}$ was counted as a success (consistent with the predicted reduced rate in open chromatin) and $K_{\text{DHS}} > K_{\text{Flank}}$ was counted as a failure. Finally, a one-tailed sign test was applied to the resulting counts to test whether rate differences between DHSs and Flanks were consistent with a reduced mutation rate at DHS sites. Similarly, to compare λ, the alternative hypothesis allowed λ to differ between DHSs and Flanks. In this case, the LRT has $df = 1$. Alignments with nominally significant differences between DHS and Flank were identified, classified and a one-tailed sign test used to assess whether differences between DHSs and Flanks in λ were consistent with the expectation of excess transition substitutions in Flanks.

To examine the influence of 5^mC on transition substitutions, a dinucleotide substitution model was used to capture context-dependent effects of CpG sites. This model used a nucleotide frequency weighted model form [85] and the HKY parameterization. The CpG transition term CG.λ was included for CpG $\leftrightarrow$ TpG and CpG $\leftrightarrow$ CpA exchanges. The CG.λ parameter measures the ratio of CpG transitions to all transitions. Note that without CG.λ, the dinucleotide model is simply the product of two independent nucleotide HKY models [85]. Because CpG is strand-symmetric and mostly methylated on both strands, CpG $\leftrightarrow$ CpA exchange also potentially arises from a methylation-associated substitution. Therefore, following the conventional notation, the term q_{ij} represents the relative rate of change from dinucleotide i to j in the instantaneous matrix $\mathbf{Q}$. The q_{ij} are defined as:

$$q_{ij,i\neq j} = \begin{cases} 0 & \text{more than one nucleotide difference} \\ \pi_x & \text{Transversion} \\ \pi_x \cdot \lambda & \text{Transition} \\ \pi_x \cdot \lambda \cdot CG.\lambda & \text{Transition involving CpG} \end{cases} \tag{3.1}$$

where π_x is the frequency of the nucleotide in dinucleotide j being substituted.

I considered two possible orders to measure CG.λ / λ heterogeneity between DHSs and Flanks. I first considered whether the rate of CpG transitions differs between DHSs and Flanks and whether the differences were independent of the mean transition effect. The null hypothesis was that DHS and Flank had the same CG.λ but different λ parameters ($CG.\lambda_{DHS} = CG.\lambda_{Flank}, \lambda_{DHS} \neq \lambda_{Flank}$); the alternative hypothesis removed the constraint on CG.λ, so that both CG.λ and λ parameters differed between DHSs and Flanks ($CG.\lambda_{DHS} \neq CG.\lambda_{Flank}, \lambda_{DHS} \neq \lambda_{Flank}$). The second path was to measure whether differences in transitions between DHSs and Flanks were independent of CpG transitions. The null hypothesis constrained λ to be equal but allowed CG.λ to differ ($CG.\lambda_{DHS} \neq CG.\lambda_{Flank}, \lambda_{DHS} = \lambda_{Flank}$). The alternative hypothesis reached the same parameterization as the first path by removing the constraint on λ with the same $df = 1$.

The Holm-Bonferroni method [86] of multiple test correction was applied to reject or accept simultaneously tested multiple hypotheses. Briefly, we assume the overall type I error rate is δ for n null hypotheses. For instance, the above analyses tested four hypotheses from both intronic and intronic regions, so $n = 8$ and we chose $\delta = 0.05$. The p-values from the n tests were first ordered and the smallest (p_0) one was compared with δ/n. If p_0 is less than δ/n, the corresponding null hypothesis was rejected and the second smallest p-value (p_1) was used to test the remaining $n - 1$ hypotheses. Thus, if p_1 is less than $\delta/(n-1)$, the null hypothesis corresponding to p_1 is rejected. This procedure was repeated until p_k is greater than $\delta/(n-k)$ where $0 \leq k < n$. Finally, the first k null hypotheses were rejected.

3.3 Results

3.3.1 DHS regions exhibit a distinct substitution rate

The samples of DHSs were carefully chosen to minimise the potential influence of natural selection on functional elements. Experimental observations indicate that the majority of DHSs that affect gene transcription are located on proximal promoters and may extend to the first exon or intron. This type of DHS is enriched in transcription factor binding sites, so they were avoided by using intergenic DHSs located far away from genes and intronic DHSs. Although some intergenic DHSs do contain remote enhancers that contribute to regulation of gene expression, the majority of intergenic DHSs have not been demonstrated to be functional. Moreover, since functional elements within DHSs are generally short DNA motifs for protein binding, choosing relatively long DHSs will lower the proportion of such sites, reducing the possibility of detecting differential substitution rates between DHSs and Flanks arising solely from natural selection operating on those functional sites.

Use of DHS regions defined in somatic tissues will make our estimates conservative. DHS regions are caused by interruption to the usual structure of nucleosome arrays, while nucleosomes at the flanking sites of DHSs tend to be particularly well-positioned [87]. For a DHS state to impact on substitution processes requires its presence in germline cells. There are two major types of DHS, constitutive and inducible. Constitutive DHSs are independent of gene expression and exist in multiple cell lines [76, 88], while inducible DHSs are induced by a number of biological factors, such as transcription factor binding, and are likely to be tissue-specific. If a DHS detected in T cells is inducible, no differences in rate or types of substitution would be expected between the corresponding DHS and Flank regions since the chromatin states in these regions are expected to be the same in germline cells. Therefore, using annotated DHSs from somatic cells increases

background noise and reduces our power to detect an influence of these chromatin states on substitution.

LRTs on substitution rate revealed that DHSs evolve significantly more slowly than Flanks in both intergenic and intronic regions. By allowing the substitution rate to differ between DHS and Flank pairs as the alternative hypothesis, 897 intergenic alignments were identified as nominally significant ($p \leq 0.05$ from LRTs) against the null hypothesis of a homogenous substitution rate. In these 897 loci, a significant majority exhibited DHSs evolving more slowly than Flanks ($K_{\text{DHS}} < K_{\text{Flank}}$, Table 3.1) which is consistent with the observation of lower efficiency of DNA repair in closed chromatin. Similar results were obtained from intronic alignments in which 878 loci were nominally significantly different and a significant proportion of these showed the predicted slower substitution rate in the DHS regions (Table 3.1).

3.3.2 DHS regions exhibit a distinct substitution type profile

In both the intergenic and intronic sequences, DHS regions exhibited a significantly lower transition substitution rate compared to the matching Flanks. It has been suggested that constraining the substitution rate in DHSs to be equal to that in Flanks may underestimate the value of λ, but the basic pattern of variation should still hold [89]. I identified 425 intergenic alignments that were nominally significant at the 0.05 level and a significant excess of these had a lower rate of transition substitutions at DHS positions (Table 3.1). Similarly, 430 intronic alignments exhibited different transitions between DHSs and Flanks. Among these intronic alignments, a significant majority showed a lower rate of transition substitutions at DHS positions (Table 3.1).

	Intronic			Intergenic		
	DHS <	DHS >		DHS <	DHS >	
Null hypothesis	Flank	Flank	p	Flank	Flank	p
$K_{\text{DHS}} = K_{\text{Flank}}$	610	268	1×10^{-31}*	639	258	2.6×10^{-38}*
$\lambda_{\text{DHS}} = \lambda_{\text{Flank}}$	238	192	0.015	252	173	7.4×10^{-5}*
$CG.\lambda_{\text{DHS}} = CG.\lambda_{\text{Flank}}$;	322	330	0.64	346	283	0.007*
$\lambda_{\text{DHS}} \neq \lambda_{\text{Flank}}$						
$\lambda_{\text{DHS}} = \lambda_{\text{Flank}}$;	225	189	0.04	223	171	0.005*
$CG.\lambda_{\text{DHS}} \neq CG.\lambda_{\text{Flank}}$						

Table 3.1: **Support for differences in total rate and substitution type profiles between DHS and Flank regions.** Null hypothesis – the hypothesis examined by the LRT; DHS < Flank (DHS > Flank) – the number of alignments that exhibited a nominally significant difference between DHS and Flank regions where the evolutionary parameter (K, λ, $CG.\lambda$) was lower (greater) in DHS than Flank; p – probability from the sign test of observing DHS rate less than Flank rate; *– significant at the 0.05 level after applying the sequential Bonferroni correction for multiple tests [86].

3.3.3 Intergenic, but not intronic, regions exhibit a distinct CpG transition rate

One possible cause of differential transition substitutions between regions is heterogeneity in CpG transitions due to 5^mC. In mammals, nucleotide methylation predominantly occurs at CpG sites and promotes C $\rightarrow$ T transition mutations. The presence of 5^mC is responsible for a considerable proportion of mutations in humans and is a dominant factor contributing to transition substitution rate heterogeneity on a localised scale in mammals [90]. Since compact chromatin has a higher density of methylated CpG dinucleotides [91], the differences in transition rates between DHSs and Flanks could be due to a greater abundance of hypermutable 5^mC in Flank sequences.

CpG transition substitutions were found to differ between DHSs and Flanks in intergenic, but not in intronic, regions. I used a dinucleotide model with a CG.λ term to represent the ratio of CpG transitions to general transitions (λ). As λ differs between DHSs and Flanks in both null and alternative hypotheses (see methods), the differential CG.λ detected here is independent of (not due to) variation in λ. In intergenic regions, 642 alignments exhibited nominally significant differences in CG.λ between DHSs and Flanks, among which a significant majority showed a lower CG.λ at DHSs ($\text{CG}.\lambda_{DHS} < \text{CG}.\lambda_{Flank}$). This observation was consistent with the conjecture that functionally-associated enrichment of 5^mC in Flanks should cause increased CpG transitions. For the intronic regions, however, there was no evidence of enrichment of $\text{CG}.\lambda_{DHS} < \text{CG}.\lambda_{Flank}$ from the 654 alignments that displayed nominally significant differences in CG.λ between DHSs and Flanks.

3.3.4 Substitutions resulting from CpG methylation do not completely account for differences in transition substitutions between DHS and Flank regions

Transition substitution differences between DHSs and Flanks remained significant in intergenic, but not in intronic, regions after accounting for CpG transitions. Since both general transitions and CpG transitions were enriched in Flank positions, the elevated CpG transition rate certainly contributed to the observed increase in transition rate. I further investigated whether the difference in λ was independent of CG.λ between DHSs and Flanks. An excess of $\lambda_{DHS} < \lambda_{Flank}$ alignments remained significant in intergenic regions. For intronic regions, it reached the nominally significant level of 0.05, but was not significant after correcting for multiple tests. Hence, variations in λ between DHSs and Flanks were not substantially affected by excluding the elevated CG.λ effect in the Flanks. Thus, the change in substitution composition between DHSs and Flanks was most pronounced for intergenic regions. Mutations of 5^mC contribute to this difference but do not entirely account for it.

3.3.5 Purifying natural selection on functional elements does not appear to be a cause of substitution heterogeneity

Purifying selection operating on functional elements within DHS sequences will contribute to a lower evolutionary rate than in Flanks. Inevitably, intergenic and intronic DHSs comprise some regulatory elements that will be under the influence of natural selection. Those critical functional sites that strongly influence phenotypes may evolve under purifying selection. Since only a handful of remote DHSs were thoroughly investigated for active protein binding sites, it was not possible for me to exclude all DNA sites under the influence of natural selec-

tion. Hence, I applied a conservative approach by eliminating alignments with annotated constrained sites to address this possibility.

Human regions that were classified as constrained elements were obtained from the UCSC databases. Using a two-state phylo-HMM to identify conserved regions from the genomic DNA, Siepel et al. (2005) [10] performed whole genome analyses for different taxa. For human, conserved elements were recognized by high conservation scores from vertebrate comparisons and are available for download from the UCSC Table browser. The segment coordinates of these elements were obtained from the table named "phastConsElements17way" in human database hg18. Constrained elements coincident with sampled DHS+Flank human sequences range from 10 to hundreds of bp in length with more than half of them less than 40 bp. Any alignments with human sequence positions annotated as conserved regions were discarded, regardless of the length of the annotation. This resulted in 2,986 intergenic and 2,936 intronic alignments respectively.

Replicating our entire analyses described above essentially achieved the same results (Table 3.2), although the statistical power was weakened due to the smaller number of loci examined. Total substitution rate was significantly lower at DHSs than Flanks for both intronic and intergenic regions. The differences in transition and CpG transition substitutions between DHSs and Flanks remained evident in intergenic, but not in intronic loci. These observations suggested that purifying selection on functional elements does not fully account for directional substitution rate heterogeneity between DHSs and Flanks.

	Intronic			Intergenic		
Null hypothesis	DHS < Flank	DHS > Flank	p	DHS < Flank	DHS > Flank	p
$K_{\text{DHS}} = K_{\text{Flank}}$	179	105	6.7×10^{-6}*	207	105	4.0×10^{-9}*
$\lambda_{\text{DHS}} = \lambda_{\text{Flank}}$	94	83	0.22	114	73	0.017*
$CG.\lambda_{\text{DHS}} = CG.\lambda_{\text{Flank}}$; $\lambda_{\text{DHS}} \neq \lambda_{\text{Flank}}$	137	130	0.34	178	106	1.2×10^{-5}*
$\lambda_{\text{DHS}} = \lambda_{\text{Flank}}$; $CG.\lambda_{\text{DHS}} \neq CG.\lambda_{\text{Flank}}$	94	75	0.08	97	79	0.10

Table 3.2: **Support for differences in total rate and substitution type profiles between DHS and Flank regions after eliminating alignments containing conserved elements.**

3.4 Discussion

These analyses have established a connection between local variation in substitution rates, both in terms of total number and types of substitution, and chromatin status at the DHS scale of 10^2 - 10^3 base pairs. A reduced total substitution rate at DHS sites was evident for intergenic and intronic regions. Differences in transition substitutions were also supported for both regions. Such differences are partially caused by elevated CpG transitions in Flank sites as (i) the magnitude of the difference decreased after considering CpG transition heterogeneity between DHSs and Flanks from both regions; and (ii) elevated CpG transition substitutions in Flank were evident in intergenic regions. However, differences in transition substitutions cannot be completely accounted for by the presence of a 5^mC effect because in intergenic regions, the differences remained significant after allowing CG.λ to vary between DHSs and Flanks.

The distinct evolutionary dynamics between DHSs and Flanks are consistent with an influence due to DNA repair. The enrichment of transition substitutions in Flank sequences supports a mutagenesis origin for substitution heterogeneity. This is because at a localised scale, replication errors should produce a homogeneous transition to transversion bias; and substitution heterogeneity arising from replication should occur on a larger scale, possibly in Mbp which is the distance between replication origins [92, 47]. For the two aspects of mutagenesis, the observed higher substitution rate in Flanks is consistent with reduced DNA repair efficiency, but contradicted by fewer lesions in closed chromatin. Therefore, DNA repair outweighs lesion formation and is the major process contributing to rate heterogeneity.

The different outcomes in transition substitutions between intergenic and intronic regions further support the differential DNA repair hypothesis. That expressed genes tend to have an open chromatin structure and a low nucleosome occupancy may cause differences in overall repair efficiency between intronic and intergenic

regions. Additionally, transcription coupled repair (TCR) only operates on transcribed regions, thus, the evolution of intronic, but not intergenic, regions are likely to be affected by TCR [93, 94]. For genes expressed in the germ line, the more accessible chromatin structure and the additional scrutiny of DNA lesions by TCR in genic regions may contribute to less significant differences in substitution rate and types between DHSs and Flanks in intronic regions. Moreover, repair of deamination products is predominantly through base excision repair (BER) pathways, which form one of the sub-pathways of TCR. Thus, TCR is a strong candidate cause for the absence of differential CG.λ and λ between DHSs and Flanks in introns.

It would seem that natural selection is not a dominant contributor to the differences in evolutionary dynamics between DHSs and Flanks. First, the results from transition and CpG transition substitutions strongly supported a mutagenic origin hypothesis as there are no known functional mechanisms operating on intergenic and intronic sequences that specifically repress transition mutations. Second, experiments have demonstrated that functional elements within DHS are mostly DNA-protein interaction motifs. As these motifs are usually short, only a small number of DNA sites within DHS are expected to undergo purifying selection. Third, cis-regulatory elements are likely to experience a high turnover rate which may lead to an accelerated substitution rate. For instance, the Encode project estimated that about half of the functional elements located in non-coding regions are unconstrained [95]. Examination of DHSs with lengths less than 300 bp also failed to detect substitution rate heterogeneity from Flanks (data not shown). As shorter DHSs harbor higher proportions of functional elements, this result further confirmed that natural selection is unlikely to account for the observed higher conservation at DHSs than Flanks. Fourth, results from alignments without constrained elements (Table 3.2) provided clear evidence of the influence of chromatin status on substitutions.

The relatively small number of significant loci found in substitution rate and

type tests may be due to factors related to data sampling. Approximately 12-13% of the alignments showed nominally significant differences in substitution rate, and fewer DHS and Flank pairs displayed significant differences in transition substitutions. Since a substantial number of DHSs are tissue specific, they do not account for substitution heterogeneity between DHSs and Flanks arising from different chromatin states. The other cause may be the choice of primate sequences. To detect substitution rate heterogeneity between different chromatin structures using comparative genomics, a prerequisite is that chromatin states are conserved among species. Hence, remotely related species, such as primates and rodents, are probably unsuitable for analyses, since chromatin states are likely to differ between them. However, the use of primates limits the statistical power from LRT, especially for the tests of transition substitutions.

This work revealed substitution rate and type heterogeneity between local open and closed chromatin structures, and established DNA repair as a likely major contributor to this heterogeneity. By using small scale regions and regions without putatively constrained elements, other factors, such as GC content and natural selection that are also associated with substitution rate heterogeneity, were excluded.

Chapter 4

Evidence That Nucleosome Placement Contributes to Localised Substitution Rate Heterogeneity

Abstract

That DHSs evolve at a slower substitution rate than compact chromatin suggests a potential role for nucleosome positioning in sequence divergence. DHSs are hypersensitive to nucleases due to the absence of nucleosomes or canonical nucleosome structures. Similarly, linkers, DNA sites between adjacent nucleosomes, are sensitive to nuclease cleavage while nucleosomal sites are protected from nuclease digestion. Thus, regular positioning of nucleosomes along chromatin would be predicted to result in local substitution rate heterogeneity that mirrors the nucleosome repeating scale. I addressed the influence of nucleosome positioning on substitution rate using ~1800 primate promoters. Phylogenetic hidden Markov model (Phylo-HMM) and phylogenetic footprinting were used to measure fluctuations in substitution rate, which were compared with nucleosome

scores from human T cells. A significant positive correlation was found between them, with up to ~50% of the variance in substitution rate accounted for. Using signal processing techniques, a dominant periodicity of ~200 bp was detected in both the spatial substitution spectrum and nucleosome scores. These results support localised variation in sequence divergence arising from a reduced rate of DNA repair associated with nucleosomes.

4.1 Motivation

Substitution rate heterogeneity between open chromatin DHS and closed Flanking sites raised the possibility that first-order chromatin compaction is a key contributor to localised variation in sequence divergence. DHSs usually extend from several hundred up to 1-2 thousand base pairs, with the minimum length corresponding to the size of a nucleosome unit. In vitro experiments revealed that DHSs induced by binding of transcription factors were probably nucleosome-free [96]. In vivo analyses of chromatin structure demonstrated that DHS were largely devoid of nucleosomes [97] or harbored nucleosomes without a canonical structure, whereas nucleosomes flanking DHS were particularly well-positioned [87]. These observations suggest that nucleosome phasing is the major difference in chromatin structure between DHSs and Flanks. Thus, nucleosome phasing is most likely to be responsible for differential substitution rates between DHSs and Flanks.

The physical structure of nucleosomes poses a natural barrier to DNA-protein interactions, thereby affecting processes such as transcription, replication, repair and recombination. A nucleosome core is formed by 145-147 bp of DNA wrapped around a histone octamer in 1.65 turns [98]. The histone octamer comprises two copies of each of the histone proteins, H2A, H2B, H3 and H4. DNA-histone contacts occur every 10.2 bp with the minor groove of DNA facing the histones. Binding of DNA to histones restricts its accessibility to other proteins. This prop-

erty was well established by nuclease treatment of genomic DNA, which digests linker sites and produces mono-nucleosomes. Consistent with this observation, functional transcription factor binding sites (TFBSs) are preferentially located at linker sites or in the DNA major groove near the border of nucleosomes so as to be exposed on the histone surface [99]. Similarly, DNA repair systems require binding of many different proteins at different stages of repair [31]. For example, the nucleotide excision repair pathway involves about 30 proteins, of which many are DNA binding proteins [100]. Therefore, the presence of nucleosomes inhibits DNA repair through reduced DNA accessibility to the repair machinery.

Experimental evidence indicates a substantial role for nucleosome positioning in DNA repair efficiency. Using a yeast mini-chromosome with well-defined nucleosome phasing, Smerdon et. al. [101, 94] showed that repair of UV-induced damage was faster in nucleosome-free and linker DNA than in nucleosome core positions. This tendency was the same for both the BER (for a review, see [64]) and NER pathways, and evident in both yeast and mammal systems [102, 103, 104, 66, 105]. It has been shown that the rate of DNA repair by human excision nuclease for nucleosomal DNA was ~10% of that for naked DNA [66]. Detailed analysis of repair processes revealed that three sequential events generally take place at nucleosomes: identification of a lesion, repair of the lesion and restoration of a functional chromatin structure. Lesions can cause structural distortion of the DNA helix that can lead to altered DNA-histone contacts and be recognised by detector proteins [61]. For BER and NER, detector proteins are able to bind to nucleosomal DNA, but with much less efficiency than to naked DNA. In the repair step, nucleosomes may be temporarily removed from the lesion site by mechanisms like chromatin remodeling (suggested by the high nuclease sensitivity of newly repaired chromatin DNA, [106, 107, 108]). The final step is reassembly of the nucleosome in its original location; a process which may require a number of enzymes (for a review, see [61]). This repair process is more complex than the equivalent repair of naked DNA, likely requiring additional energy. It therefore

suggests repair of lesions in nucleosomal DNA will be less efficient.

Detailed analyses of the efficiency of nucleosomal site repair suggested that DNA repair rate is site-dependent. DNA repair rate heterogeneity not only exists between linkers and nucleosomes, but also within nucleosomes. An examination of the DNA repair efficiency of UV-induced lesions in the yeast *URA3* gene revealed that the repair rate was slowest at the nucleosome centre and gradually increased towards the periphery [109]. Reduced repair efficiency at central nucleosomal sites increases the longevity of DNA lesions that occur there, increasing the likelihood of their conversion to mutations during the next round of DNA replication. Consequently, site-dependent DNA repair rate heterogeneity at nucleosomes and linkers predicts site-dependent substitution rate heterogeneity if nucleosome organization is conserved among species.

Examinations of sequence divergence from various sources support a causal relationship between evolutionary rate and nucleosome positioning. The total substitution rate was lower in linker sites than nucleosomal sites in intergenic regions in yeast [11], as well as in exons [14]. Detailed analysis of substitution rate at each site relative to nucleosome dyads supported a site-dependent hypothesis [13]. Such an association further predicts a periodic pattern in the spatial distribution of substitution rate because nucleosome units repeat every 200±40 bp throughout eukaryotic genomes [110]. This was evident downstream of TSSs in fish where the evolutionary rate exhibited $\sim$ 200 bp periodicity and was concordant with nucleosome occupancy [15]. In the human, single nucleotide polymorphism (SNP) density also displayed a 200 bp periodicity around TSSs in CpG island promoters [111].

While the observations reported by these studies are consistent, they raise further questions about the origin of substitution rate heterogeneity and the influence of individual nucleosomes. Two conflicting interpretations have been proposed regarding an origin from natural selection or mutation processes. Washietl et al.

[13] found that substitution rate heterogeneity between linker and nucleosomal sites in yeast was independent of genic or intergenic locations, nucleotide positions within a codon, and dinucleotide frequencies. Thus, they concluded that natural selection on functional elements was not responsible for this rate heterogeneity. In contrast, Warnecke et al. [14] suggested natural selection was responsible because codons favoring nucleosome binding were overrepresented in nucleosome associated positions, but underrepresented in linker-associated codons. In both these studies, the estimation of substitution rate was from concatenation of DNA sites from disjoint genomic locations with the same classification of chromatin states. Thus, the influence of individual nucleosomes on the spatial distribution of substitution rate has not been addressed.

With the development of high throughput techniques, genome-wide nucleosome mappings are now available for many species. Briefly, in these experiments, genomic DNA is digested with an appropriate amount of micrococcal nuclease (MNase) with the result that only intact nucleosomes remain. The resulting mononucleosomes are then (optionally) purified by a selected histone antibody and DNA fragments with a length of ~150 bp isolated from an agarose gel. The collected DNA is then either hybridised to probes on a microarray or sequenced using next-generation sequencing to determine their genomic locations. Both techniques have been applied to the genomes of yeast [11, 99, 112], worm [113] and human [114, 115, 116]. These data revealed important associations between nucleosome organization and biological functions (such as long nucleosome-free regions upstream of TSS in yeast), and make extensive investigation of the influence of nucleosomes on substitution rate possible.

In this chapter, the relationship between positioning of individual nucleosomes and the spatial distribution of sequence divergence is examined. Localised substitution rate heterogeneity was detected and compared with nucleosome density signals on promoters. Hypotheses that were tested include: (i) that localised substitution rate heterogeneity exists in promoter sequences; (ii) that the spatial

distribution of substitution rate is correlated with nucleosome density signals; (iii) that the spatial substitution spectrum shows periodic patterns; (iv) that periodic patterns detected from spatial substitution spectra and nucleosome density signals are concordant. I found support for all these hypotheses.

4.2 Methods

4.2.1 Promoter data with nucleosome annotations

We obtained nucleosome positions on human promoters from the published study of Ozsolak et. al [12]. The genomic coordinates of nucleosomes were downloaded from the Gene Expression Omnibus (GEO) under the accession number GSE6385. These coordinates were mapped to the UCSC human database hg17 which was based on the NCBI35 assembly. To be consistent with Ensembl release 50, hg17 coordinates were converted to those of the NCBI36 assembly using the UCSC LiftOver tool [84]. This resulted in 3,555 promoter regions (each ~1500 bp long) with 37,991 nucleosome positions from 7 cell lines. Since repeat sequences were excluded from microarray probes, these nucleosome positions do not cover repeats on promoters.

Promoter alignments of human, chimpanzee and macaque were obtained from Ensembl release 50 based on human coordinates. Genes with annotated nucleosome positions on their promoters were identified using Ensembl gene annotations. Genes with their TSS within 3 Kbp of other protein-coding genes were excluded. Alignments were obtained for genomic regions extending from the annotated human gene TSS to 1,500 bp upstream. Alignments with less than 1,000 columns after removal of columns containing gaps or N characters were excluded. This resulted in 1,849 promoter alignments. Note that this sample contained genes with annotated CpG islands.

4.2.2 Chip-seq nucleosome signals

Genome-wide nucleosome mapping from high-throughput sequencing was from Schones et al. [115], whose procedure is restated here for completeness. Nucleosome density was represented by nucleosome scores with a sliding window of 10 bp. Nucleosome scores were calculated by counting the number of sequencing tags within 80 bp upstream on the '+' strand and 80 bp downstream on the '−' strand. A higher nucleosome score indicates a higher probability of nucleosome occupancy. Both activated and resting T cell nucleosome scores are available from `http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellnucleosomes.aspx`. We used the nucleosome scores from resting T cells since activated T cells were treated with antibodies to stimulate an immune response in the study of Schones el al. [115] while resting T cells were closer to a natural state.

4.2.3 A Phylogenetic hidden Markov model to measure spatial substitution rate heterogeneity

A phylo-HMM is a combination of a standard site-independent phylogenetic model and a hidden Markov model (HMM) that allows the substitution rate to change from one site to the next and allows autocorrelation of substitution rate among neighboring sites. The likelihood calculation for a given alignment is the same as that used in the phylogenetic models described in Chapter One. If X_i represents the ith column in the alignment, the probability of X under an evolutionary model ψ can be defined as a function of four parameters: $P(X_i|\psi) = P(X_i|\mathbf{Q}, \tau, \beta, \pi)$, where $\mathbf{Q}$ is the substitution rate matrix, τ is a tree topology, β is a vector of branch lengths for τ, and π is a vector of equilibrium nucleotide frequencies. Substitution rate variation among sites is achieved by scaling the vector β by a factor r. Gamma distributed rate variation has been found to provide a good fit in various data sets [117, 118]. In this case, a discrete

gamma distribution with n categories is used to derive r [119]. To be computationally efficient, the probabilities of sites belonging to each rate category are set to be equal. Therefore, the likelihood under a rate heterogeneity model can be written as:

$$P(X_i|\psi) = \sum_{j=1}^{n} \frac{1}{n} \cdot P(X_i|\mathbf{Q}, \tau, r_j\beta, \pi) \tag{4.1}$$

The autocorrelation of substitution rates among sites is achieved by assuming Markov dependence of rates at adjacent sites with a transition parameter bin switch (ε). For example, if there are two state ($n = 2$) categories f, s and alignment column X_{i-1} belongs to category f, the probabilities that column X_i belongs to category f or s are $1 - \varepsilon$ and ε respectively. The likelihood with transition parameter ε of an alignment is then computed using a dynamic-programming algorithm [120] and the parameter values can be maximised using standard numerical optimisation routines. Having maximised a model, the posterior probabilities of site categories are determined by posterior decoding [120] which can then be used for classification.

We used a phylo-HMM to formally test the existence of spatial substitution rate heterogeneity. The base phylogenetic model was a HKY substitution model. The null model was defined as two equiprobable substitution rate categories, designated as fast and slow, with gamma distributed rate-heterogeneity. The alternative hypothesis allowed non-independence of site rate class with the additional parameter ε. LRTs were used to find evidence of clustering of fast / slow rate category sites. The difference in degrees of freedom was 1. To correctly maintain the transition status among sites, alignment gaps were kept. The estimated posterior probabilities of sites belonging to the fast category (p_{fast}) were used as an indicator of substitution rate variation. p_{fast} values corresponding to human sequence residues were extracted from the alignment and used to generate Figure 4.1.

4.2.4 Phylogenetic footprinting to measure spatial substitution rate heterogeneity

Maximum likelihood phylogenetic footprinting [121] (hereafter, footprinting), which has different assumptions to the phylo-HMM, was also applied to measure substitution rate variation along sequences. Footprinting is a sliding window method that fits a model to sequential, overlapping windows. Parameters estimated from each window can then be plotted against the window's position and thus tracked along the sequence.

In our case, we designed the footprinting algorithm to be comparable to the phylo-HMM model. The HKY model was first fit to the entire alignment. Substitution rates were then measured for 100bp windows which were moved progressively down the alignment in 5 bp steps. To be consistent with a phylo-HMM that incorporates a single λ for the entire alignment, the value of λ in the HKY model for each window was constrained to that estimated from the full alignment. The standard PyCogent optimisation routines were applied to maximise the likelihood of the model. The substitution rate K was calculated as the sum of branch lengths of the unrooted phylogenetic tree comprising human, chimpanzee and macaque. Finally, K was assigned to the middle position of the window.

Data sampling for the footprinting was further distinguished from the phylo-hMM by the treatment of gap columns in the alignment. For the phylo-HMM, alignment columns are correlated by the state transition parameter ε, so gap columns were retained to maintain correct relations to ancient states. By contrast, the footprinting approach assumes independence of alignment columns. Thus, deleting gap columns, a common approach to measure the pure substitution process, will not affect the likelihood of other columns. For phylo-footprinting, I deleted alignment columns with gaps in the human sequence so that the windows represented the same number of human nucleotides.

4.2.5 Statistical testing of the correlation between substitution spectrum and nucleosome score

A bootstrap procedure was employed to test the correlation between K generated from phylo-footprinting and nucleosome score signals from Chip-seq experiments. Both the nucleosome score and K statistic series are not independent. For instance, any single alignment window used for estimation of K includes 95% of the sites from the adjacent window. As this causes statistics estimated from the windows to be non-independent, standard significance testing of the correlation coefficient is not appropriate. Instead, a bootstrap process using randomised blocks [122] was applied to estimate the probability that $\rho \neq 0$. The block length in use was equal to the length of the footprinting window size. Specifically, randomised series of K, denoted as K_{rand}, were generated by drawing, with replacement, blocks of data from the observed K until the length of K_{rand} series equalled that of the observed K. If a random draw required a series that exceeded the length of the data, the draw was continued from the beginning. Correlation coefficients (ρ_{rand}) were computed from K_{rand} and the observed data (ρ). This process was repeated 2000 times, generating a distribution of ρ_{rand}. The probability (p) of observing a larger ρ by chance was computed by the frequency of $\rho_{rand} \geq \rho$. Finally, a multiple test correction [86] was performed to find promoters with experiment-wide significant p values. Because of the large number of promoters being considered, only loci for which no single $\rho_{rand} \geq \rho$ were identified. In contrast, for negatively correlated loci, all $\rho_{rand} > \rho$.

4.2.6 Signal period estimation

Regular spacing of nucleosomes predicts an oscillation in evolutionary rate with low rate positions located in the linker regions between nucleosomes. As illustrated by Sasaki et. al [15], substitution rate peaks occur every ~200 bp down-

stream of the TSS in the fish genome. Whether this is also a general feature of nucleosome organization on promoters in human has not been addressed. This periodicity is detectable using appropriate signal processing techniques.

Periodic patterns in both substitution spectrum and nucleosome score were measured by a Discrete Fourier Transform (DFT) method. DFT has been extensively used in periodicity analyses for genomic signals. For a numerical signal such as K whose periodicity is vaguely sinusoidal, DFT is appropriate. According to its conventional definition, the DFT is written as:

$$X[f] = \sum_{n=0}^{N-1} x[n] \exp\left(-j\frac{2\pi nf}{N}\right), f = 0, 1, ..., N-1, \tag{4.2}$$

where N is the signal length, $x[n]$ corresponds to K for the nth alignment window, f is the discrete frequency index corresponding to a period p $= N/f$. If the signal contains a single dominant periodicity, it can be estimated from the magnitude spectrum of $|X[f]|$ using the maximum likelihood estimator

$$\hat{f} = \arg\max_{f<\frac{N}{2}} |X[f]| \tag{4.3}$$

The period resolution drops dramatically with the increase of estimated period. From the definition of the DFT, the period interval between two adjacent periods $\triangle p = \frac{N}{f} - \frac{N}{f+1} = \frac{N}{f(f+1)}$ where $f = 1, 2, ..., N/2$. This clearly shows that the smaller the frequency index f (longer period) is, the larger $\triangle p$ is. For example, footprinting used a window size of 100 bp and a step of 5 bp for 1500 bp long alignments. N from footprinting is therefore 280 ((1500-100)/5). Because the interval of N points is 5 bps, the period from DFT is 1400, 700, ... ,233, 200, 175, 155, ..., 40, 38.9, 37.8, ... ($p = \frac{N}{f.5}$). Thus, long periods such as 200 and 210 cannot be distinguished by DFT with the same resolution as shorter periods such as 39 and 40. Since the periods of interest are about the size of nucleosomes, some estimates of the confidence of the estimated period were required.

To understand the limitations of period estimation, the signal (footprinting or nucleosome score) was modeled as a sinusoid plus noise and the Cramer-Rao bound (CRB) was computed as the variance of the estimated $\hat{f}$ and $\hat{p}$. For the frequency estimator $\hat{f}$, the CRB was solved by Tretter [123] and has been investigated extensively. Following the same assumptions, such as a single sinusoid of amplitude A with white noise of variance σ_w^2, the CRB for the period estimator $\hat{p}$ has recently been shown by Epps et al. (Epps J, Ying H and Huttley GA unpublished data) to be

$$\text{var}(\hat{p}) \geq \frac{6\sigma_w^2}{A^2 N^3} \left(\frac{p^2}{2\pi^2} \right)^2 \tag{4.4}$$

which shows that the variance of the estimator $\hat{p}$ is strongly determined by the period length p, the inverse of the signal-to-noise ratio $SNR = \frac{A^2}{\sigma_w^2}$ and the signal length N. Retaining the assumption of a single (dominant) sinusoid in additive noise, the signal-to-noise ratio is written as (Epps J, Ying H and Huttley GA unpublished data):

$$\hat{SNR} = \frac{\sum_{f=0}^{N/2} |S[f]|^2}{\sum_{f=0}^{N/2} ||X[f]| - |S[f]||^2} \tag{4.5}$$

where

$$S[f] = X[\hat{f}] \frac{\sin(\pi(f - \hat{f}))}{\sin(\pi(f - \hat{f})/N)} \tag{4.6}$$

4.3 Results

4.3.1 The substitution rate was significantly heterogeneous along promoter sequences

A two-state phylo-HMM was employed to address the influence of individual nucleosomes on substitution rate. Two states were used for a number of reasons.

First, I was interested in distinguishing the influence of two chromatin states, namely nucleosomal and linker sites. Due to differences in DNA repair at these sites, nucleosomal DNA is expected to correspond to the fast state and linker DNA to the slow state. Second, the phylo-HMM implementation in PyCogent is limited to 2 states. Therefore, the scaling factors (r) for the fast and slow substitution states were greater than 1 and less than 1 respectively. All model parameters were estimated directly from the alignments.

A considerable number of promoters showed significant spatial substitution rate heterogeneity. LRTs were performed to formally test the existence of spatial clustering of fast and slowly evolving DNA sites. Among 1,849 individual promoter alignments, 505 was nominally significant, which was substantially higher than expected by chance ($1849 \times 0.05 \approx 92$). After correcting for multiple tests [86], 37 alignments remained significant. This analysis suggested that substitution rate heterogeneity does exist on many promoters, subject to the caveat that the existence of spatial variation in sequence composition was not addressed by the phylo-HMM.

4.3.2 Phylo-footprinting displayed similar substitution rate heterogeneity

Footprinting was also applied to find evidence for a spatial distribution of the substitution process. To reduce the homogenising effect on the substitution signal of windows spanning nucleosome and linker sites, a window size of 100 bp was used. With constrained λ for each footprinting window, footprinting measures the spatial distribution of substitution rate denoted as the substitution spectrum. We expect this should be correlated with the phylo-HMM p_{fast} where a higher probability means a higher substitution rate and vice versa. As illustrated for the *CDX2* and *FGF5* promoters (Figure 4.1), the substitution spectra from the phylo-HMM and phylo-footprinting did resemble each other with high Pearson's

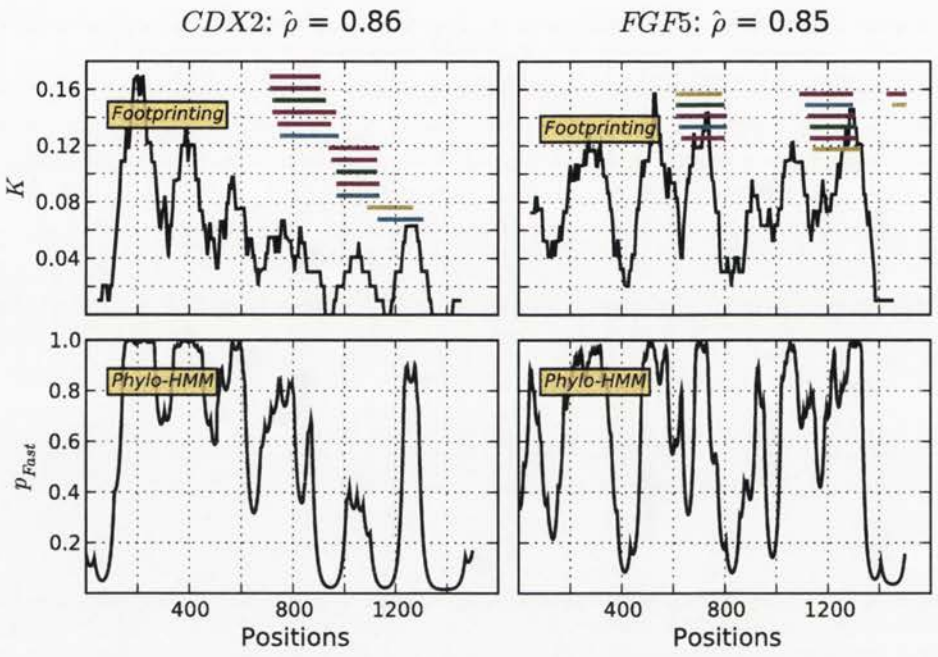


Figure 4.1: **Comparison of the spatial substitution rate variation estimated from phylogenetic footprinting and a phylo-HMM.** The upper panel row shows the substitution spectra from footprinting, measured as the the sum of tree branch lengths (K), from the promoters of *CDX2* and *FGF5*. Each horizontal line indicates a nucleosome inferred from one of the seven cell lines where: magenta represents one of the four tumour cell lines: A375, T47D, MCF7 and MALME; green represents the IMR90 cell line, cyan represents the PM cell lines; and yellow represents the MEC cell line [12]. The lower panel shows the posterior probabilities of a site classified as 'fast' (p_{fast}), estimated from the phylo-HMM. $\hat{\rho}$ is the estimated Pearson's correlation coefficient of the footprinting and phylo-HMM signals.

correlation coefficients (ρ). Among 505 significant promoters from phylo-HMM, ~44% showed strong correlations ($\rho > 0.5$) between the spatial distribution of K and p_{fast}.

Substitution spectra from footprinting were used for the subsequent analyses. The disadvantage of phylo-HMM is that it assumes homogeneous sequence composition across the alignment, an assumption clearly violated for promoters containing CpG islands. While the same assumption is made in footprinting, it is over a much smaller spatial scale, making violation less likely. Furthermore, for

those alignments with p values greater than 0.05 from the phylo-HMM LRT, the substitution spectra were unstable, which means that a different set of posterior probabilities may occur with a second run of the phylo-HMM. In contrast, substitution spectra from phylo-footprinting were robust. Therefore, all substitution spectra from phylo-footprinting were available to be compared with experimental nucleosome signals.

4.3.3 The spatial substitution spectra and nucleosome scores are significantly correlated for some loci

Based on observations from our DHS analyses and other studies [13, 15, 14], nucleosomal sites were expected to evolve at a higher substitution rate than linkers. Substitution rate differences between DHSs and Flanks implied that DNA regions with well-positioned nucleosomes would evolve faster than nucleosome-depleted regions. Examining the substitution rate at each nucleotide within nucleosomes revealed that the substitution rate is site-dependent with the highest rate at the nucleosome dyad and with lower rates towards the linker regions. These results predict that nucleosomes locate at peaks (high substitution rate regions) of substitution spectra. When substitution spectra are compared with the nucleosome score signals (where higher score represents higher nucleosome density), we expect a positive correlation between them.

The correlation between substitution rates and discrete genomic regions annotated as nucleosome locations was ambiguous. Substitution spectra were initially compared with nucleosome positions defined from microarray experiments [12]. A comparison without formal statistical tests was carried out due to the different data types from the two measurements: numerical signals from substitution spectra, but genomic regions from nucleosome annotations. Nucleosome regions located to mixed positions on the substitution spectra. For instance, on the *BLNK* promoter, nucleosomes were mostly located at peaks of K, while on the

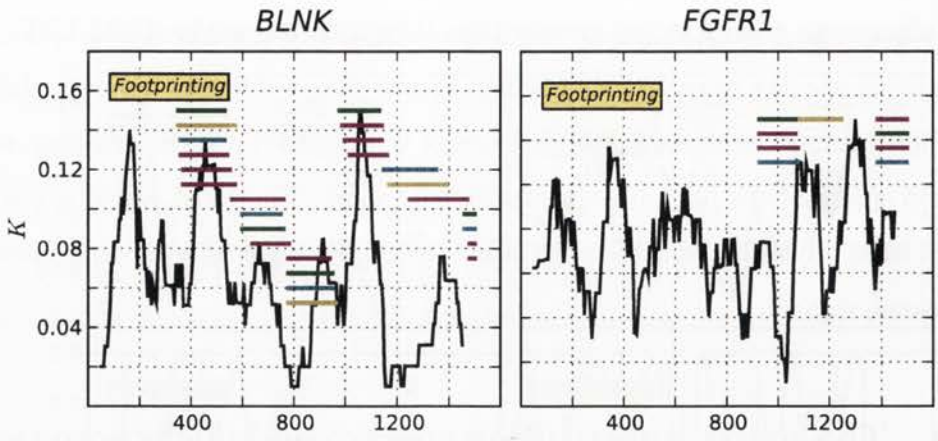


Figure 4.2: **Comparison of the substitution spectra estimated from footprinting with nucleosome annotations.** Substitution spectra were estimated from footprinting for the *BLNK* (left) and *FGFR1* (right) promoters respectively. Nucleosome positions from 7 cell lines are indicated in the same way as in Figure 4.1

FGFR1 promoter, nucleosomes were located in troughs (Figure 4.2). Since annotated nucleosome positions are sporadically distributed along the sequence with different coverage among promoters, they were considered unsuitable for formal testing. During the period of this work, more nucleosome data with numerical signals became available which made a formal comparison possible.

Substitution rate was significantly correlated with a continuously distributed measure of nucleosome location. We identified 125 nominally significant ($p < 0.05$) loci with positive correlations among 1,793 promoters for which nucleosome score data were available. 5 out of 125 were significant after correcting for multiple tests (Table 4.1). The two most positively correlated loci are shown in Figure 4.4.

Negative correlations arose from substantially out-of-phase signals, which appeared to arise from attributes of both data types. A quantile-quantile plot of the distribution of the probabilities from bootstrap tests against the quantiles from the uniform distributions displayed departure at both ends, but predominantly at $\rho < 0$ (Figure 4.3). 276 promoters displayed nominally significant negative ρ

($p > 0.95$), 15 of which remained significant after correcting for multiple tests. The two most negatively correlated loci are shown in Figure 4.4. An excessive number of loci with negative correlations suggest oscillations exist in both signals but they are out-of-phase. However, an assessment of genomic features on the promoters identified other potential causes for the negative correlations. On the *PRELP* promoter (Figure 4.4), for example, nucleosome scores were missing in a long repeat region. This arises because repetitive sequence reads are likely to match multiple positions in the genome and are thus discarded before calculation of nucleosome scores [115]. Nucleosome scores on CpG islands and DHS regions are also sometimes missing or very low. Since DHS regions are likely to be tissue-specific while substitution rate only measures mutations arising in the germ line, exact measurements from the substitution process can be discordant with the annotations.

As the limitations of the nucleosome score data preclude direct comparison, we considered an indirect approach to assess the relationship between substitution spectra and nucleosome mapping signals. A comparison of K with the nucleosome positioning signals showed that a number of promoters exhibited the predicted significant positive correlation, but an excess of loci with significant negative correlation also existed. This observation prevents us from making strong conclusions about the generality of the correlation between substitution rate heterogeneity and individual nucleosomes. Given the sensitivity of high-throughput sequencing and microarray techniques to repetitive sequences and differences in nucleosome phasing between somatic cells and germline cells, complex relationships are not unexpected as the evaluation heavily relied on the accuracy of somatic nucleosome positions. I therefore applied an alternative approach that examines the generality of the effect of nucleosomes on the substitution processes, but does not depend on knowing the coordinates of individual nucleosomes.

Symbol	$\hat{\rho}$	p_{boot}
PRELP	-0.6740	1.0000
EYA1	-0.6682	1.0000
GIP	-0.6675	1.0000
GNAI2	-0.6517	1.0000
DAP	-0.6345	1.0000
PTP4A1	-0.6301	1.0000
PYGL	-0.6245	1.0000
GRAP2	-0.5757	1.0000
RHBDL2	-0.5755	1.0000
PPP5C	-0.5620	1.0000
ATG7	-0.5490	1.0000
TMEM103	-0.5485	1.0000
SMOX	-0.5188	1.0000
PRPF19	-0.5109	1.0000
MFGE8	-0.5017	1.0000
RGN	0.5899	0.0000
CACNA1G	0.6255	0.0000
PCDH8	0.6623	0.0000
GSTO1	0.6721	0.0000
ARG1	0.6762	0.0000

Table 4.1: **Promoters with correlated K and nucleosome scores** Promoters with correlated K and nucleosome scores. ρ is the Pearson's correlation coefficient from K and the nucleosome score. p_{boot} is the probability that ρ is not equal to 0, estimated using a bootstrap procedure with 2000 replicates. The listed promoters were significant after multiple test correction.

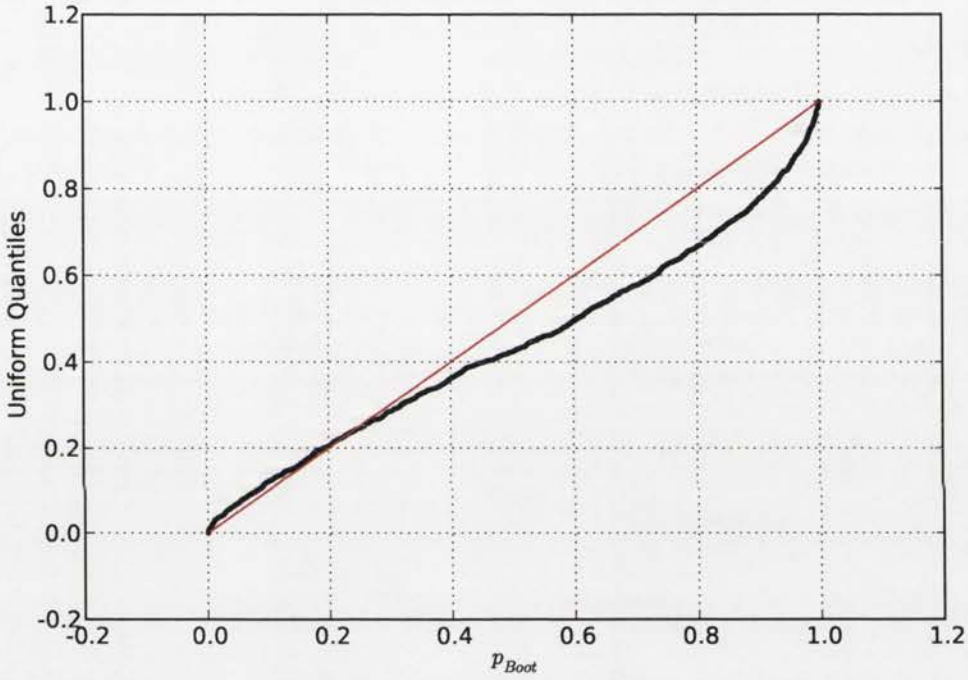


Figure 4.3: **Quantile-Quantile plot of the bootstrap probability distribution against the uniform distribution.** Departures between the probability of bootstrap tests and the uniform distribution were observed at both ends. The red line represents the expected relationship when the null hypothesis: no correlation between K and nucleosome score, is true.

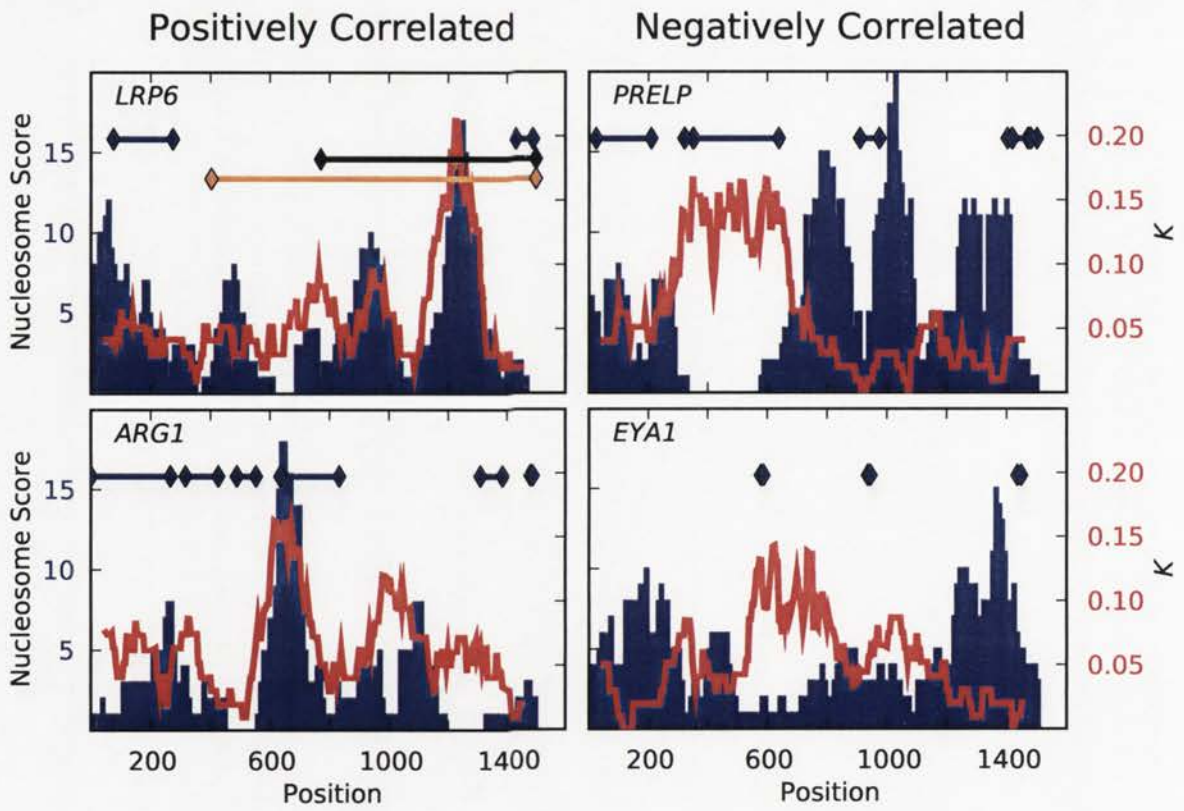


Figure 4.4: **Comparison of the substitution spectrum with nucleosome score.** Example genes exhibiting a positive correlation are shown in the left column, a negative correlation in the right column. x-axis is alignment position; y-axis with black label is the nucleosome score with data shown as a histogram; y-axis with red label is the estimated K from footprinting with data shown as a red line. Black, orange and blue horizontal lines with diamond marks at the ends represent CpG islands, DHS sites and repeat sequences respectively.

4.3.4 An $\sim$ 200 bp oscillation in both substitution rate and nucleosome score

The beads-on-string model of nucleosome phasing along a genomic sequence suggests the substitution rate will oscillate if rate heterogeneity arises from nucleosome positioning. The consistent sequence span of the nucleosome repeating unit suggests that they will occur at regular intervals across the genome producing a periodic pattern measurable by signal processing techniques. Given the evidence reported above for the influence of chromatin structure on substitution processes, the substitution spectra should likewise exhibit a periodicity of the size of the nucleosome plus linker. Since linker sites range from 10 to 80 base pairs in mammals, the expected periodicity is 200$\pm$40 bps [110].

The application of DFT analysis and associated measurements to determine dominant periods is illustrated by the analyses of two promoters, *DUSP* and *FZD2* (Figure 4.5). A DFT transforms substitution rate series into period series with the signal power (amplitude) corresponding to the strength of each period. Periods from DFTs were measured in nucleotides. Periodic components of the substitution spectrum appear as peaks in the amplitude spectrum. For instance, in the *FZD2* promoter there were three dominant peaks from DFT power series corresponding to 700, 233, and 175 bp respectively. These three periods were candidates for the periodicity in the substitution spectrum of the *FZD2* promoter. To better understand the limitations of the estimated periods, a CRB (interpreted as the variance of the estimated period) threshold of 0.2 was employed. Estimated periods with a CRB greater than 0.2 were excluded. After applying the threshold, the periods with the greatest and penultimate power were designated as the main and secondary periods respectively. For *FZD2*, the period of 700 bp had a CRB of 0.649. Examination of the substitution spectrum indicates this period likely derives from the large amplitude peaks of K at the alignment positions of $\sim$ 100-300, $\sim$ 700-1000, and $\sim$ 1300-1400. Since a 700 bp period from

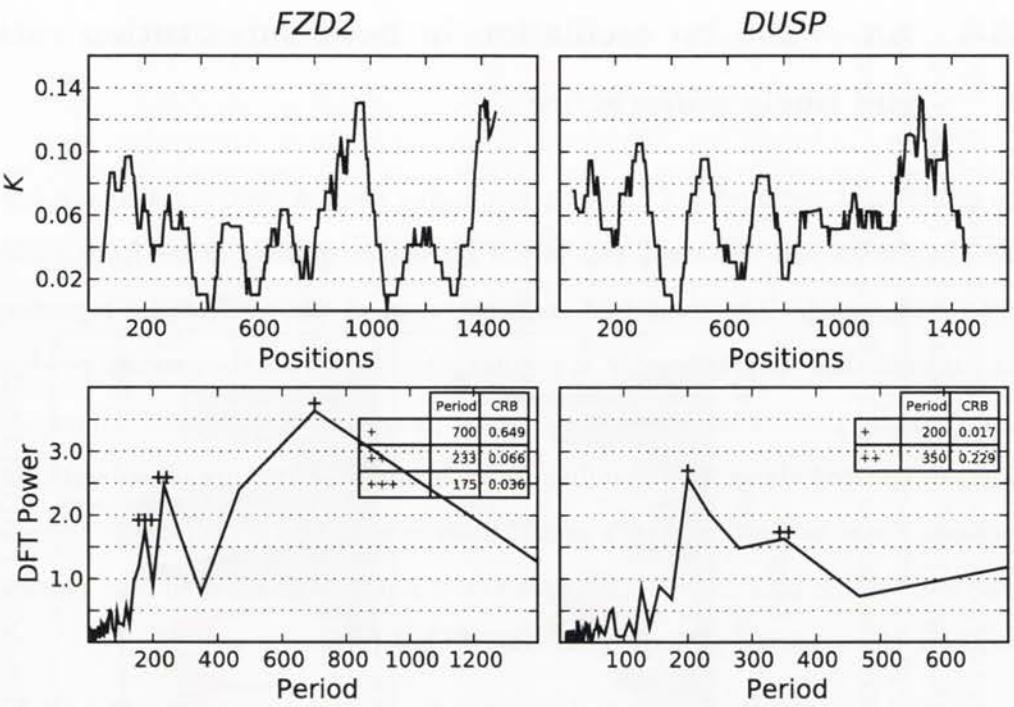


Figure 4.5: **Signal analysis of substitution spectra by DFTs.** The left and right columns correspond to the DUSP and FZD2 promoters respectively. The upper row shows K while the lower row is DFT. Periods of the substitution spectrum appear as peaks in the DFT spectrum. The 1st, 2nd and 3rd highest peaks are marked with a corresponding number of '+'s, and their period lengths and CRBs are shown in the tables.

a 1400 bp signal can only be repeated twice, the high CRB value indicates a lack of confidence. Thus, the 700 bp period was not considered the dominant periodic component in K. The next two peaks of DFT power, 233 bp and 175 bp, both exhibited a CRB < 0.07. These two peaks were then designated as the main and secondary periods for *FZD2* respectively. For the *DUSP* promoter, only a single period of 200 bp had a CRB < 0.20, so it was designated the main period.

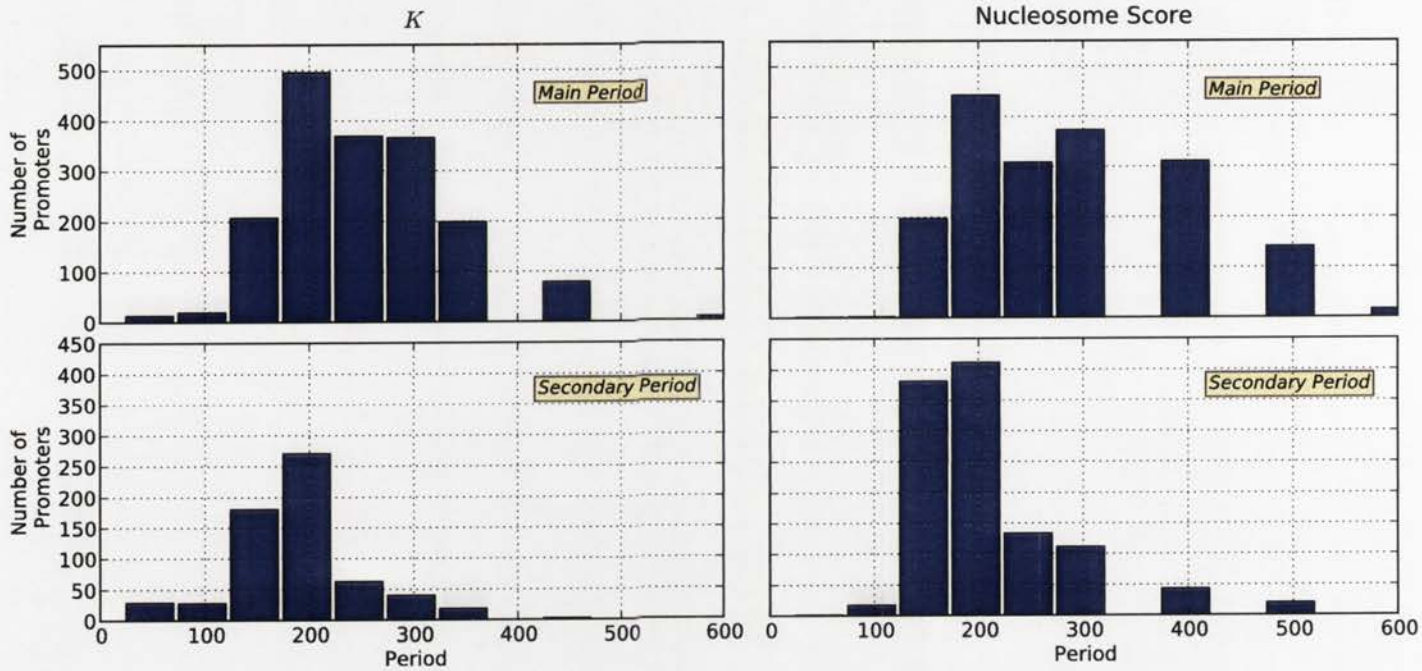


Figure 4.6: **Periodicity estimated from substitution spectra and nucleosome scores.** Both K and raw nucleosome scores exhibit an $\sim$ 200 bp period in primate promoters. Upper and lower rows show the frequency histograms of the periods classified as Main and Secondary respectively after eliminating periods with CRB$\geq$0.20. The left and right columns display the periods from K and nucleosome scores respectively.

A dominant period of $\sim$ 200 bp in substitution spectra is evident across all promoters, consistent with an influence of nucleosomes. The main and secondary periods (Figure 4.6) evaluated from K were mainly distributed between 125 and 375 bp, with the mode being the 200 bp bin that spanned periods from 175 to 225 bp. We compared these results to periods measured from nucleosome scores derived from a Chip-seq study [115] on the same promoter data set. Again, the most frequent period in both main and secondary periods (Figure 4.6) was $\sim$ 200 bp. The consistency in periods between K and the experimental data further support an influence by nucleosome phasing on substitution rate heterogeneity.

Periods estimated from substitution spectra and nucleosome scores were concordant regardless of the choice of CRB threshold. Since longer periods repeat less often for a fixed alignment length, larger CRB are expected. To avoid arbitrary filtering of certain periods, we repeated the above analyses of period distributions with additional CRB thresholds of 0.50, 0.10 and 0.05 (Figure 4.7). As expected, the periods selected tend to be smaller, and the number of observed main and secondary periods decreased as the CRB threshold was reduced. However, the distributions of periods estimated from K and the nucleosome score were generally similar for all CRB thresholds.

4.4 Discussion

These analyses established a general correlation between localised substitution rate heterogeneity and nucleosome placement. Localised substitution rate heterogeneity on promoters was consistently identified by both the phylo-HMM and footprinting approaches. Comparison of the spatial substitution spectra with experimental data of nucleosome density signals revealed both positive and negative correlations. Although a direct effect of nucleosome phasing on substitution rate cannot rely on somatic cell-derived nucleosome positions, numerous significant correlated loci suggested that both signals oscillate along the sequences. Using

signal processing approaches, a dominant $\sim$ 200 bp periodicity was evident from both signals. That substitution spectra and nucleosome density data share similar periodic components strongly supports an influence of nucleosome positioning on substitution processes.

Localised variation in substitutions on promoters support a mutation driven hypothesis. The conflicting explanations of a mutagenic [13] or a selective origin [14] for substitution rate heterogeneity between linkers and nucleosomes indicate that confounding evolutionary forces co-exist in most genomic sequences in yeast. Because the yeast genome is compact with $\sim$ 70% of the genome encoding proteins, intergenic regions are generally only hundreds of base pairs long. This region may be strongly selective as the proximal regions of both 5' and 3' ends of genes are generally more conserved than putatively neutral sequences. The use of protein-coding sequences [14] makes it difficult to distinguish natural selection on nucleosomal codon usage from spatial patterning of selection along a gene (e.g. [58]). By contrast, the proportion of sequence under *purifying* selection in vertebrate genomes is likely to be much smaller ($\sim$ 3-8%). Comparison of core promoters with adjacent four-fold degenerate sites has revealed promoters as faster evolving in primates [124]. Hence, only a small number of sites in the sample of long promoter sequences are expected to be under purifying selection. Moreover, the periodic pattern of substitution spectra is readily explained by the recurrent placement of nucleosomes given the experimental evidence of an influence of nucleosomes on mutation processes.

Both significant positive and negative correlations between nucleosome positions and the spatial substitution spectra illustrate the challenges in identifying nucleosome footprints from substitution spectra in multi-cellular organisms. Many factors, including methodological artifacts and biological activities, can affect the accuracy of each of the statistics. Besides the challenge of repetitive and/or low complexity sequences mentioned before, the other methodological issue affecting estimation of nucleosome scores involves the relatively low-coverage from

genome-wide Chip-seq data. A peak-detection algorithm [116], which successfully detected nucleosomes from multiple histone modification data sets [125], failed to find well-positioned nucleosomes from nucleosome score signals from most promoters (data not shown). The phylo-footprinting approach is not affected by repeat sequences, but is affected by low-complexity sequences such as CpG islands. CpG islands, which are enriched in GC nucleotides, make multiple alignment challenging and potentially lead to under-estimation of evolutionary divergence. This may not be a problem using the closely related primate species since oscillations in K were clear within CpG islands (Figure 4.4). Another issue affecting efforts to correlate substitution spectra with nucleosome positions is the assumption that chromatin status is conserved among species. This may be violated in some circumstances, thus affecting estimation of K. We suggest, however, that the most critical confounding factors are biological activities. The existence of distinct soma and germline cell lineages in multi-cellular organisms raises the issue of complex nucleosome phasing arrangements in vivo. This could be a cause of the failure to detect site-dependent substitution rates within nucleosomal DNA in primates by Washietl et al (2008) [13]. Examining nucleosome positions from multiple cell lines revealed substantial changes in nucleosome organization on human promoters (Thomas O, Tremethick D and Huttley GA, unpublished data). As heritable mutations are restricted to those from germline cells while the nucleosome data used were derived from somatic tissues, ambiguous correlation between substitution spectra and nucleosome positioning signals is not unexpected.

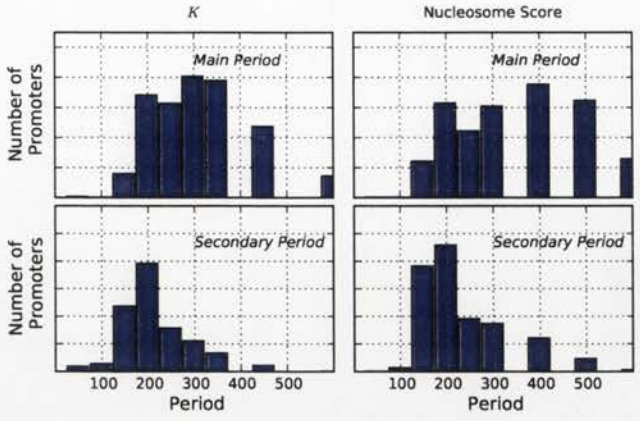
An $\sim$ 200 bp periodicity in both the substitution spectra and nucleosome scores is consistent with an effect of nucleosome placement on localised rate heterogeneity. The above-mentioned limitations of measuring direct correlation between the two independent signals motivated us to extend our assessment by addressing the general prediction that the nucleosome repeating unit along sequences will cause an oscillating pattern. The advantage of this approach was that it did

not require knowledge of the genomic coordinates of nucleosome phasing in the primate germ line. The disadvantage was that it is an indirect assessment of the role of nucleosomes in the spatial distribution of K. Since the exact nucleosome intervals vary, I compared the distribution of primary and secondary periods obtained from K and nucleosome scores from matching promoters. A dominant $\sim$ 200 bp periodicity from the spatial distribution of K itself suggested an influence of nucleosome positioning; and the concordance in periodicity with nucleosome scores further supported this interpretation. This correspondence was robust to the choice of maximum variance in the period estimation.

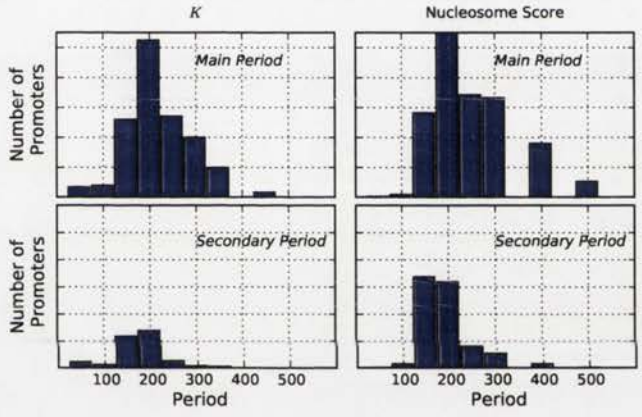
Multiple periodicities in both substitution spectra and nucleosome scores from many promoters indicate the complexity in nucleosome positioning. If nucleosomes are spaced at constant intervals, a single dominant period is expected from nucleosome scores and substitution spectra. However, multiple periodic components exist in most promoters. Many factors affect nucleosome organization on promoters. In yeast, there is a long nucleosome-free region upstream of most TSSs. In humans, a nucleosome-depleted region seems evident from both microarray and Chip-seq signals, especially on expressed gene promoters. Since open chromatin structures are often found in promoters, such as DHSs and CpG islands, promoters may exhibit unusual nucleosome organization in some of their sequences. Another factor relates to nucleosome activities, e.g. chromatin remodeling or nucleosome sliding, which have been associated with "fuzzy" nucleosomes. Fuzzy nucleosomes occupy DNA sequences longer than the standard 147 bp and are estimated to comprise about half of the nucleosomes in the yeast genome. All these features will affect nucleosome positioning in vivo and thus influence the estimation of periodicity from Chip-seq and substitution signals.

This work revealed the existence of localised substitution rate heterogeneity on promoters and established a general correlation between the substitution process and individual nucleosome phasing. This association can be accounted for by differential DNA repair between nucleosomes and linkers. Although a clear cor-

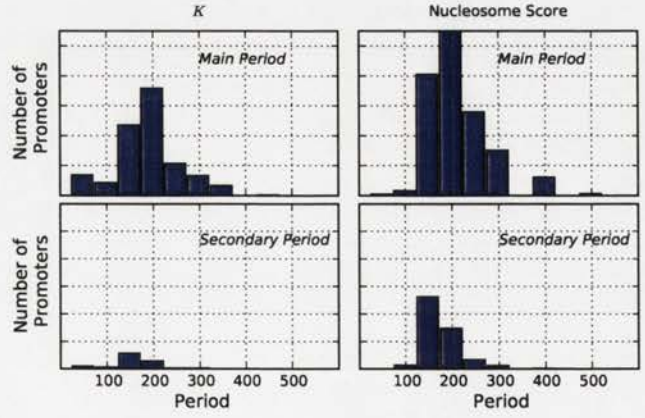
relation of nucleosome positioning on the spatial distribution of substitutions was not achieved using current somatic nucleosome position data, it suggests that it is possible to predict nucleosome placement through comparative genome analyses when more accurate nucleosome positioning data become available.



(a) CRB threshold = 0.50



(b) CRB threshold = 0.10



(c) CRB threshold = 0.05

Figure 4.7: **Period distributions from K and nucleosome scores were consistent under different CRB thresholds**

Chapter 5

The Impact of DNA Methylation on Protein Coding Sequence Evolution

Abstract

The modified nucleotide 5^mC presents a striking contrast in shaping the evolution of genomic sequences in vertebrates. When the great mutability of 5^mC encounters natural selection within coding sequences, the interplay between these two evolutionary forces predicts a distinct mutation-selection balance at CpG sites, where 5^mC predominantly occurs. We contrasted evolutionary dynamics between genomes that methylate and do not methylate their DNA, using primates and yeast respectively. We further took advantage of a well-characterized framework for measuring the mode of natural selection in protein-coding genes. We applied codon substitution models with parameters that measure mutation properties and selective strength affecting CpG-containing codons. From genome-wide analyses, a greatly increased CpG transition substitution rate was evident for the majority of primate genes, while only a small proportion of yeast genes displayed a modestly elevated CpG transition rate. CpG-encoded amino acids were found

to experience stronger purifying selection only in primates relative to the same set of amino acid exchanges from non-CpG mutational events. These results are consistent with a shifted mutation-selection balance at CpG-containing codons in primates, putatively arising from 5^mC. Furthermore, we examined the association between CpG effect and disease. We observed significant enrichment of disease-associated variation in genes with elevated CpG transition rates and stronger purifying selection on CpG codons, indicating stronger functional significance of CpG-encoded amino acids in primates.

5.1 Motivation

Population genetics theory [126] establishes that the equilibrium frequency of an allele is governed by an inverse relationship between mutagenicity and natural selection. To illustrate this, consider a genetic locus with two alleles A and a. If the mutation rate from A to a is μ and from a to A is ν, the equilibrium frequencies are $\nu/(\mu+\nu)$ for allele A and $\mu/(\mu+\nu)$ for allele a. If the two alleles exert different effects on phenotype, the equilibrium frequencies will depend on the mode of natural selection. Suppose allele a is recessive, strongly deleterious and present at a low frequency (so the mutation rate ν is ignored); the relative fitnesses of each genotype AA, Aa, and aa are 1, 1, and 1-s respectively where s is the selection coefficient against the aa genotype. The equilibrium frequency of allele a is then approximately $\sqrt{\frac{\mu}{s}}$ [126]. In this simple scenario, higher mutability (μ is large) or less disadvantage (s is small) of allele a will both increase the equilibrium frequency of allele a. This within-population effect should also be manifest between species. Sequence divergence between species should exhibit the same pattern because substitutions are fixed genetic variants. The fixation probability of a genetic variant is governed by natural selection in that it is higher for beneficial mutations but lower for harmful mutations. Consequently, at a particular sequence position, the functional significance is expected to be

inverse proportional to the probability of a sequence state being substituted. This relationship suggests an opportunity to exploit the differences in the mutability of codons to identify functionally significant positions.

The characteristic position-specific amino acid profiles of individual proteins reflect the distinct functional roles of each residue and represent the set of exchangeable amino acids that can be accommodated without impairing protein function. The naturally occurring 20 amino acids differ greatly in their physicochemical properties, which determine the interchangeability between amino acids. For instance, Isoleucine (I) and Leucine (L), which are both hydrophobic, are largely interchangeable, while Alanine (A) and Proline (P) are not. Moreover, whether a specific position can tolerate amino acid exchanges is also determined by the specific position in a protein. To illustrate this, we consider the Hox gene clusters that encode homeodomain transcription factors and comprise hundreds of proteins from insects to mammals. The homeodomains (HD) of DNA binding sequences are highly conserved within orthologues and between paralogues. However, the linker region that connects the HD to the the hexapeptide (HX) domain displays great variation [127, 128] in primary sequence. Thus, the homeodomain is selectively constrained by amino acid physico-chemical properties to maintain the integrity of molecular function, while the linker region is highly tolerant, suggesting the exact identify of the amino acid at these positions is less critical and hence these positions are more likely to be "neutral". This variability is well illustrated by the sequence logos (e.g. `http://pfam.sanger.ac.uk/family?entry=PF00046#tabview=tab3`). Consequently, while the evidence that nonsynonymous substitution rates are typically lower than synonymous substitution rates indicates that a substantial fraction of amino acid substitutions are deleterious, some positions exhibit a high degree of plasticity with multiple amino acids tolerated. Therefore, the relative frequencies of the different amino acids that make up proteins will depend on both the selection intensity and mutation processes operating on the corresponding set of

codons.

The dominant point mutation process in vertebrates involves CpG-encoded codons. As described in Chapter One, CpG sites exhibit higher mutability than other dinucleotides due to DNA methylation on cytosine (5^mC). This mutation pressure is applicable to CpG-containing codons since exons are generally methylated [129, 130]. Consistent with this expectation, estimates from empirical rate matrices suggested that substitutions for CpG-encoded amino acids are more permissive than the average (Huttley GA unpublished data). On the other hand, CpG-containing codons encode a collection of amino acids with different physicochemical properties that are constrained by natural selection. Thus, the reasoning above concerning mutation-selection balance predicts that the presence or absence of a CpG codon at a specific protein position will be dependent on the mode of natural selection.

We consider three scenarios of selection for destroying a CpG-encoded codon and compare methylated CpG equilibrium frequencies with presumably non-methylated ancestors. Given that the mutation rate from CpG → TpG / CpA is about one order of magnitude higher than the reverse, (i) CpG codons at neutrally evolving positions will move towards a lower frequency of CpG; (ii) codons where CpG mutations are only slightly deleterious will also evolve to a low CpG frequency if selection cannot effectively oppose the mutation pressure; and (iii) strong functionally significant CpG codons will be maintained by substantial purifying selection against CpG loss. Thus, neutral or nearly neutral CpGs have been gradually destroyed leading to their serious underrepresentation in vertebrate genomes [131, 132]. A direct consequence of this process is that well-preserved CpG sites indicate an increased likelihood of functional significance, and their relative abundance among all CpG sites will increase with the reduction of neutral CpG sites. This conjecture has an important implication for finding amino acids that affect protein function and are associated with disease.

Previous examinations of the interaction between CpG mutability and natural selection within protein-coding sequences have been flawed. One approach was to apply codon substitution models [133, 134] as these allowed formal hypothesis testing under the phylogeny-based maximum likelihood framework. Extending Goldman and Yang's model [133], Huttley [135] introduced CpG-context dependent substitution parameters to assess relative substitution rate and selective strength at CpGs compared to other types of nucleotide substitutions. CpG-containing codons were estimated to exhibit strikingly higher substitution rates and a lower ratio of nonsynonymous to synonymous substitution rates in the *BRCA1* gene. However, the baseline model of Goldman and Yang was subsequently found to be unsuitable for measuring context-dependent processes [85, 136]. Moreover, the amino acids encoded by CpG may be subjected to different selective constraints than other amino acids due to their distinct physicochemical properties. This possibility was not explicitly addressed. Schmidt et. al. (2008) [137] compared mutation probabilities arising from CpG and non-CpG content for the same amino acid exchanges using a parsimony-based method. They found a substantially elevated CpG nonsynonymous transition substitution rate by comparing the same amino acid exchanges with or without CpG context, and reduced fixation probability for CpG nonsynonymous transitions by comparing the ratios of transition rates within a CpG context to that of outside a CpG context between nonsynonymous and other substitutions. However, no formal tests for statistical significance were conducted and background selective constraints affecting CpG-encoded amino acids were not adjusted for. Furthermore, since closely related primates were used, concatenating substitutions from a large number of genes was required. It has been well established that the substitution rate varies substantially across mammalian genomes [47, 138, 3, 139, 140, 4]. Thus, using a CpG transition rate derived from whole genomes is not reasonable. Finally, as CpG codons occupy various positions in different proteins, selective constraints operating on CpG codons will differ from gene to gene. The analyses

of Schmidt et. al [137] did not identify individual proteins that contained CpG codons subjected to strong negative selection.

In this chapter, we modify the codon model approach of Huttley [135] to use a robust model form and improve the estimation of the mode of natural selection affecting CpG-containing codons. The objectives were to assess the extent of elevated CpG mutation properties within coding sequences and whether CpG-encoded amino acids were subjected to greater purifying selection than an appropriately defined background rate. Using yeast as biological controls whose genomes are putatively methylation free, we show that codon substitution models can formally establish specific context-dependent mutation profiles and distinguish unique selective constraints on a subset of defined amino acids. We further present evidence of stronger selective constraints on CpG codons by examining disease-causing genes.

5.2 Material and Methods

5.2.1 Statistical models of codon evolution

We employed the CNF (Conditional Nucleotide Frequency) matrix model form [136] to measure codon evolution. For a continuous-time Markov substitution process, the instantaneous rate of substituting codon i by codon j has the general form:

$$q_{ij,i\neq j} = \begin{cases} 0 & \text{more than one nucleotide difference} \\ \pi_x \cdot r(i,j) & \text{otherwise} \end{cases} \tag{5.1}$$

where π_x is the equilibrium frequency, and $r(i,j)$ is the product of rate parameters affecting exchanges between the codons. For instance, in the standard codon model form, $r(i,j)$ includes combinations of the parameters ω and λ for: synonymous transversions $[r(i,j) = 1]$; synonymous transitions $[r(i,j) = \lambda]$; nonsynony-

mous transversions $[r(i,j) = \omega]$; and, nonsynonymous transitions $[r(i,j) = \lambda \cdot \omega]$. The CNF form differs from the codon substitution models of Muse and Gaut (MG94 [134]) and Goldman and Yang (GY94, [133]) in the definition of π_x. The MG94 model uses the frequency of the nucleotide in codon j that differs from codon i, with the result that the equilibrium codon frequencies are the product of nucleotide frequencies (normalized for the omission of stop codons). This multiplicative feature of the MG94 model is unlikely to be satisfied in coding sequences and has been shown to bias parameter estimates when codon frequencies are not multiplicative [136]. π_x in the GY94 model is the frequency of codon j, so the equilibrium codon frequencies readily match those observed, but this formulation confounds the single nucleotide substitution event with the frequency of other sequence states. This confounding has the undesired effect of causing the GY94 model to show context-dependent effects when they do not exist [85]. Therefore, parameter estimates in both the MG94 and GY94 models can be biased by sequence composition. π_x in the CNF model, by contrast, is the frequency of the nucleotide in codon j that differs from codon i, conditional on the other two nucleotides in codon j, expressed as:

$$\pi_x = \begin{cases} \pi_1 | j_2, j_3 & i_1 \neq j_1,\ i_2 \doteq j_2,\ i_3 \doteq j_3 \\ \pi_2 | j_1, j_3 & i_1 \doteq j_1,\ i_2 \neq j_2,\ i_3 \doteq j_3 \\ \pi_3 | j_1, j_2 & i_1 \doteq j_1,\ i_2 \doteq j_2,\ i_3 \neq j_3 \end{cases} \tag{5.2}$$

where i_1, i_2, i_3 and j_1, j_2, j_3 represent the nucleotide states at the three codon positions in codon i and j respectively. The merit of the CNF model is that it nests the independent substitution process, but also allows equilibrium codon frequencies to be non-multiplicative (see [136] for a complete explanation).

We used a CNF model incorporating general time-reversible (GTR) terms based on the previous demonstration of the robustness of parameter estimates of this form [136]. The 6 parameters within the GTR (Table 5.1) model represent all

possible unique, reversible, exchanges between nucleotides [141]: $r_{A\leftrightarrow C}$, $r_{A\leftrightarrow G}$, $r_{A\leftrightarrow T}$, $r_{C\leftrightarrow G}$, $r_{C\leftrightarrow T}$, and $r_{G\leftrightarrow T}$. This set of parameters will be subsequently referred to as r_{GTR}. These parameters have been applied to the codon substitution model previously [142, 136]. Hence, with GTR terms, $r(i,j)$ is defined as:

$$r(i,j) = \begin{cases} r_{GTR} & \text{synonymous substitutions} \\ r_{GTR} \cdot \omega & \text{nonsynonymous substitutions} \end{cases} \tag{5.3}$$

where ω represents the average selective strength operating on all codons. This is the baseline model for subsequent analyses.

To measure CpG-related substitution dynamics, a rate parameter for exchanges occurring within a CpG context were included in the model. Following the notation of Huttley [135], the parameter G (Table 5.1) measures the relative rate common to all CpG substitutions (transversions and transitions) and parameter $G.K$ (Table 5.1) measures the relative rate common to all CpG transitions (CpG $\leftrightarrow$ TpG; CpG $\leftrightarrow$ CpA). If repair of a T/G mismatch is frequently accompanied by a complete replacement of a mismatched nucleotide, a single additional term G will be adequate to measure the methyl-CpG mutation property. Since CpG is strand symmetric and mostly methylated on both strands, replacements at both positions are potentially methylation-induced. This results in the following definition of $r(i,j)$ as:

$$r(i,j) = \begin{cases} r_{GTR} & \text{synonymous substitutions} \\ r_{GTR} \cdot G & \text{synonymous substitutions involving CpG} \\ r_{GTR} \cdot \omega & \text{nonsynonymous substitutions} \\ r_{GTR} \cdot G \cdot \omega & \text{nonsynonymous substitutions involving CpG} \end{cases} \tag{5.4}$$

Hence, G greater than 1 (less than 1) means an elevated (reduced) substitution

Table 5.1: **Substitution model terms**

Term(s)	Definition
GTR	General, time-reversible nucleotide substitution parameters projected into a codon model. It contains five independent parameters: $r_{A\leftrightarrow C}$, $r_{A\leftrightarrow G}$, $r_{A\leftrightarrow T}$, $r_{C\leftrightarrow G}$, and $r_{C\leftrightarrow T}$, whereas $r_{G\leftrightarrow T}$ is constrained to equal 1
ω	Ratio of nonsynonymous to synonymous substitution rate
α	Ratio of nonsynonymous substitution rate involving amino acid exchanges affected by CpG transitions to general nonsynonymous substitution rate
G	Ratio of CpG substitution rate to corresponding nucleotide substitution rate
$G.K$	Ratio of CpG transition substitution rate to corresponding nucleotide transition substitution rate
$G.K.\omega$	Ratio of nonsynonymous CpG transition substitution rate to corresponding nonsynonymous nucleotide transition substitution rate

rate at CpG sites compared to the model background. On the other hand, if T/G mismatch repair is faithful, modeling CpG transitions will be adequate to measure the methyl-CpG mutation property alone with $r(i,j)$ defined as:

$$r(i,j) = \begin{cases} r_{GTR} & \text{synonymous substitutions} \\ r_{GTR} \cdot G.K & \text{synonymous transitions involving CpG} \\ r_{GTR} \cdot \omega & \text{nonsynonymous substitutions} \\ r_{GTR} \cdot G.K \cdot \omega & \text{nonsynonymous transitions involving CpG} \end{cases} \tag{5.5}$$

In addition, CpG transitions may be distinguished from the CpG transversion rate, in which $r(i,j)$ is defined as:

$$r(i,j) = \begin{cases} r_{GTR} & \text{synonymous substitutions} \\ r_{GTR} \cdot G & \text{synonymous transversions involving CpG} \\ r_{GTR} \cdot G \cdot G.K & \text{synonymous transitions involving CpG} \\ r_{GTR} \cdot \omega & \text{nonsynonymous substitutions} \\ r_{GTR} \cdot G \cdot \omega & \text{nonsynonymous transversions involving CpG} \\ r_{GTR} \cdot G \cdot G.K \cdot \omega & \text{nonsynonymous transitions involving CpG} \end{cases} \tag{5.6}$$

Under the influence of methylation, $G.K$ is predicted to be significantly greater than 1.

The influence of natural selection on CpG transitions was measured by a parameter corresponding to the interaction of CpG transitions and nonsynonymous substitutions. The existing parameter ω represents the common influence of natural selection on the rate of substitution for all amino acids. The amino acid exchanges, resulting from CpG transitions, may also result from non-CpG events (Table 5.2). For instance, substitutions between alanine and valine can arise from CpG-containing codons (e.g. GCG $\leftrightarrow$ GTG) or not from CpG-containing codons (e.g. GCA $\leftrightarrow$ GTA). The rate of replacements between the 12 methylation-affected amino acids (hereafter MAA, Table 5.2) are likely to differ from the "average" amino acid exchange due to the distinctive changes in physico-chemical properties. To assess whether the positions at which the MAAs are CpG-encoded evolve differently from other MAA positions, we first introduced a parameter α (Table 5.1) to represent the nonsynonymous exchange rate common to all MAA positions. We then introduced parameter $G.K.\omega$ (Table 5.1) as previously defined [135], assigned specifically to nonsynonymous transition changes within CpGs, to

Parameter		Amino acid exchange	codon substitution
α	Nonsynonymous CpG transition ($G.K.\omega$)	A↔V	GCG↔GTG
		C↔R	TGC↔CGC; TGT↔CGT
		H↔R	CAC↔CGC; CAT↔CGT
		L↔P	CTG↔CCG
		L↔S	TTG↔TCG
		M↔T	ATG↔ACG
		Q↔R	CAA↔CGA; CAG ↔CGG
		R↔W	CGG↔TGG
	Nonsynonymous non-CpG substitution	A↔V	GCA↔GTA; GCC↔GTC; GCT↔GTT
		L↔P	CTA↔CCA; CTC↔CCC; CTT↔CCT
		L↔S	TTA↔TCA
		R↔W	AGG↔TGG

Table 5.2: **Codon substitutions represented by α and $G.K.\omega$** Amino acids exchange arising from CpG transitions were represented by parameter $G.K.\omega$. The same set of amino acid exchanges arising from both CpG transitions and non-CpG substitutions were represented by parameter α.

assess the effect of selection on CpG-encoded MAAs. Since $G.K.\omega$ is defined for a subset of α, if distinct selection constraints operate on CpG sites, $G.K.\omega$ will be different from α. Thus, $r(i,j)$ from the model with the richest parameterization

is defined as:

$$r(i,j) = \begin{cases} r_{GTR} & \text{synonymous substitutions} \\ r_{GTR} \cdot G.K & \text{synonymous transitions involving CpG} \\ r_{GTR} \cdot \omega & \text{nonsynonymous substitutions} \\ r_{GTR} \cdot \alpha \cdot \omega & \text{nonsynonymous substitutions involving MAA at non-CpG events} \\ r_{GTR} \cdot \alpha \cdot G.K \cdot G.K.\omega \cdot \omega & \text{nonsynonymous transitions involving CpG} \end{cases} \quad (5.7)$$

5.2.2 Hypothesis Testing

Hypotheses were tested by LRTs for each parameterization as described in Chapter One. We tested the support for parameters using hierarchical hypothesis testing. The modeling approach was to initially evaluate the statistical support of individual context-dependent terms, followed by joint models. There were alternative orders of fitting for joint models and both were considered, as discussed below.

Modeling Notation

The model notations introduced here are succinct expressions for different models and will be used throughout this chapter. The basic codon substitution model is the CNF model, combining equations 5.1 and 5.3. The baseline CNF model with the additional term G is represented as CNF+G. Use of the '+' symbol does not mean the terms are added in calculating the maximum likelihood, but represents the inclusion of the terms in the parameterization. Moreover, expression of additional terms in a different order results in an equivalent model. For example, CNF+G+G.K is the same as CNF+G.K+G. The model with the richest

parameterization in this chapter is CNF+G.K+α+G.K.ω.

Modeling Paths

We first defined the baseline model as a CNF model incorporating GTR and ω terms (Table 5.1). This model provides background nucleotide substitution rates and selective constraints for all codons.

We then evaluated CpG-specific substitution properties from the CpG context-dependent parameter G and $G.K$ (Table 5.1). Although the prior biochemical evidence strongly implicates an elevated CpG transition rate, we considered the possibility that repair may be imprecise. If there was a general elevation of CpG substitution affecting both transitions and transversions, the $G.K$ term could be significant by itself against the CNF baseline because it captures some of the general effect. If only CpG transitions were elevated, a G term could be significant by itself because it includes transitions. Accordingly, both possible orders of fitting G and $G.K$ must be considered (Figure 5.1). For path I, we added the G term first and then the $G.K$ term. This resulted in *probabilities* p_1 and p_2 which were from LRTs comparing CNF+G versus CNF and CNF+G+G.K versus CNF+G, respectively. Since $G.K$ defines a subset of the substitutions of G, the latter distinguishes CpG transitions from CpG transversions. Additionally, p_1 has incorporated a CpG transition effect which will reduce the statistical power of p_2. Conversely, path II assessed $G.K$ first, followed by the G term that produced *probabilities* p_3 and p_4, respectively.

We finally assessed the strength of natural selection specifically operating on CpG transitions through the parameter $G.K.\omega$ (Figure 5.1B). As previously, distinctive selection constraints operating on non-CpG MAA were considered by alternative orders of fitting α and $G.K.\omega$. Because we have a clear prior expectation regarding the CpG transition effect, the null model was the CNF+G.K model from genes with $p_3 < 0.05$ (the results presented below support this expectation).

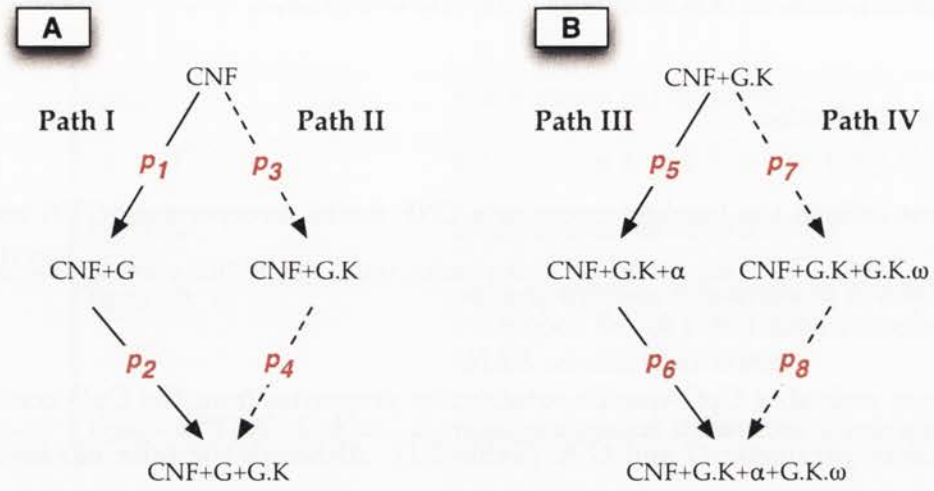


Figure 5.1: **Flow diagram illustrating four paths for nested model parameterizations and LR tests**. Arrows represent the direction of the LRTs and $p_1 - p_8$ represent the corresponding probabilities. (A) Testing for G and $G.K$ terms with two alternative ways to achieve joint models with both G and $G.K$ terms. (B) Testing for $G.K.\omega$ and α terms. It follows from the first step of path II with two alternative ways to achieve a joint model with both $G.K.\omega$ and α terms.

Support for distinctive selective constraints common to the MAAs, represented by parameter α, were evaluated by LRTs that compared CNF+G.K+α versus CNF+G.K (p_5, Figure 5.1). Subsequently, whether CpG-encoded amino acids exhibited different selective constraints from non-CpG MAAs was assessed by a LRT of CNF+G.K+α+G.K.ω versus CNF+G.K+α (p_6, Figure 5.1). Alternatively, we first considered the support for $G.K.\omega$ which measures the selective constraints on CpG codons compared with all the other codons. The corresponding LRT compared CNF+G.K+G.K.ω versus CNF+G.K and produced *probability* p_7. Finally, we added the α term mainly to evaluate the selective constraints on MAA from non-CpG events. This was assessed by *probability* p_8 from the LRT comparing CNF+G.K+G.K.ω+α versus CNF+G.K+G.K.ω.

5.2.3 Data sampling

Ensembl release 54 was used to obtain human single nucleotide polymorphism (SNP) data, human sequences, human genes and their orthologs in other primate genomes.

Primate protein-coding sequences

Human nuclear protein-coding genes that have orthologs in chimpanzee, orangutan, and macaque genomes were sampled (Figure 5.2). To avoid ambiguity from multiple gene families, only orthologous genes defined as having a one-to-one relationship were considered. For each gene with orthologs in the three primates, the longest coding sequence (without the terminal stop codon) among transcripts were collected for each species. Furthermore, only coding sequences that could be translated were retained.

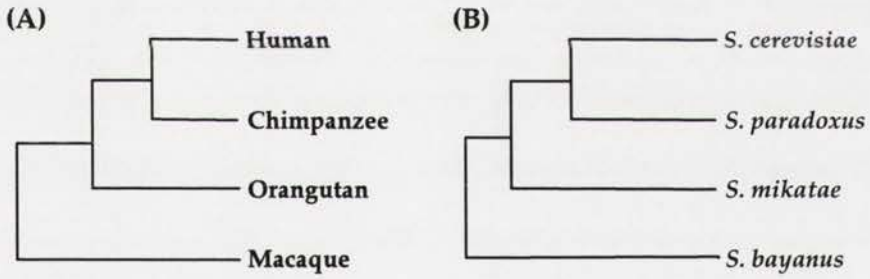


Figure 5.2: **Phylogenetic tree for (A) primates and (B) yeast**

Yeast protein-coding sequences

Yeast nuclear protein-coding genes were sampled from four closely related species, namely *Saccharomyces cerevisiae, S. paradoxus, S. mikatae, and S. bayanus* (Figure 5.2). Orthologous yeast sequences were downloaded from the Saccharomyces Genome Database (SGD) ftp site at `ftp://genome-ftp.stanford.edu/pub/`

yeast/sequence/fungal_genomes/Multiple_species_align/other/MIT_Spar_Sbay_Smik_Scer. Each downloaded file is an alignment from the four yeast species corresponding to one *S. cerevisiae* ORF and its flanking sequences provided by Kellis et al. [143]. In the source files, exon sequences were distinguished by capital letters, while non-coding sequences, including intergenic and intronic sequences, were written in lower case. To be consistent with the workflow for primates, yeast coding sequences were extracted and aligned using the PyCogent codon aligner (see below). Additionally, the same criteria for filtering primate coding sequences were also applied.

Coding sequence alignments

A progressive pair-hidden Markov model (HMM)-based multiple sequence aligner was applied to align coding sequences for each gene from both primate and yeast. This algorithm was initially developed by Loytynoja and Goldman [144] and later implemented in PyCogent [58]. The PyCogent aligner takes an arbitrary substitution model to compute the probability of match states. This approach provides a better solution than traditional alignment algorithms that consider insertion the same as deletion, resulting in overmatching of sequences but underestimation of insertions [144]. In PyCogent, a substitution model is required to perform alignment, while a phylogeny tree and parameter values are optional. If the phylogenetic tree is not specified, a neighbor-joining tree is determined from pairwise distances estimated with the same substitution model.

We used a codon substitution model to align primate coding sequences. The codon aligner allows incorporation of nucleotide and nonsynonymous substitution rate parameters and ensures that indels preserve the reading frame. The baseline CNF model described above was used. Alignments generated by the codon aligner were further filtered such that any codon columns containing 'N' or '-' characters were eliminated. If the remaining columns were less than 200 codons ($<$ 600

bp), the alignment was discarded. This resulted in 10,044 primate and 1,934 yeast coding sequence alignments respectively.

OMIM genes

The NCBI Online Mendelian Inheritance in Man (OMIM) database (`http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim`) was used to query disease-associated genes. OMIM catalogues human genes and genetic disorders with evidence from the literature [145]. Each gene record provides information like gene description, gene function, and gene structure. Some genes have allelic variants, which were carefully selected according to the criteria of (i) the first allelic variants to be discovered, (ii) high frequency in a population, (iii) mutations leading to different disorders, and so on. Note that not all allelic variants in OMIM cause disease, e.g. the CCR5 gene record includes alleles that confer resistance to HIV infection. However, since most alleles are related to pathological disorders, we used genes with allelic variants to represent disease-associated genes.

OMIM accession numbers and associated symbols that corresponded to genes were downloaded on 29 July, 2009. OMIM gene symbols were then used to query the Ensembl database to find corresponding human genes in Ensembl. If not found, alternative symbols from the OMIM gene table (available on `http://www.ncbi.nlm.nih.gov/Omim/Index/genetable.html`) were employed. This resulted in 11,697 human nuclear protein-coding genes recorded in both the Ensembl and OMIM databases. Among these genes, 6,699 genes were included in my sampled primate alignments, with 1,434 genes classified as disease-associated.

Human Variation data

Human biallelic nuclear protein-coding gene SNPs were obtained from the Ensembl variation database. These SNPs were mainly imported from NCBI dbSNP, but also included some from other sources like the supporting databases for the

Affy GeneChip 100k Array. Both validated and non-validated SNPs were considered. Validated SNPs were those that have been genotyped for a certain number of individuals within a population. A total of 173,180 SNPs were extracted among which 100,453 were nonsynonymous and 72,727 were synonymous. Each SNP was further classified by substitution types according to its allelic string. A transition allele is represented by the strings A/G, G/A, C/T or T/C; and others are regarded as transversion alleles. The 5' and 3' flanking sequences of a SNP were also acquired to determine whether it was within a CpG context. For example, a SNP with flanking sequences of 5'-C A/G G-3' is a CpG allele, while 5'-T A/G G-3' is not.

5.3 Results

We used yeast coding sequences as a biological negative control as the sampled yeast genomes are considered substantially methylation-free [146, 132]. For primates, we selected two genes, *BRCA1* and *F8*, to illustrate the estimated maximum log likelihood and parameter values from each codon substitution model. *BRCA1* was modeled by Huttley [135] under the Y98 model [54]. Comparing parameter values estimated from these two models would allow us to assess whether the conclusions made previously still hold. The *F8* gene was among the first to be identified with CpG-associated disease-causing mutations. Defects in the *F8* gene are associated with haemophilia A and CpG transition allelic variants contribute to 25% of overall nucleotide substitutions and 48% of recurrent alleles [147]. These observations suggest distinct functional encoding by CpG-containing codons in the *F8* gene.

Model	**lnL**	$\hat{\omega}$	$\hat{\alpha}$	$\hat{G}$	$G\hat{.}K$	$G.\hat{K}.\omega$	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
CNF	-8798.66	0.57												
CNF+G	-8765.00	0.58		7.05			0.00*							
CNF+G.K	-8767.86	0.54			8.87				0.00*					
CNF+G+G.K	-8762.84	0.56		3.97	2.29			0.04*		0.00*				
CNF+G.K+α	-8766.41	0.60	0.59		12.20						0.09			
CNF+G.K+G.K.ω	-8767.61	0.57			10.99	0.72							0.48	
CNF+G.K+G.K.$\omega + \alpha$	-8766.37	0.60	0.55		11.47	1.18						0.78		0.12

Table 5.3: **Statistics from analyses of primate *BRCA1* using the CNF baseline model**. Model – substitution models applied to estimate parameter values; lnL – log-likelihood; $p_1 - p_8$ – probabilities from LRTs, see Figure 5.1 for corresponding null and alternative hypotheses; * – nominally significant at the 0.05 level. See Table 5.1 for definition of terms.

Model	**lnL**	$\hat{\omega}$	$\hat{\alpha}$	$\hat{G}$	$\hat{G.K}$	$G.\hat{K}.\omega$	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
CNF	-7685.43	0.47												
CNF+G	-7677.11	0.45		3.42			0.00*							
CNF+G.K	-7677.94	0.44			4.15				0.00*					
CNF+G+G.K	-7676.54	0.44		2.37	1.76			0.28		0.10				
CNF+G.K+α	-7674.77	0.52	0.34		8.25						0.01*			
CNF+G.K+G.K.ω	-7673.03	0.51			12.14	0.16							0.00*	
CNF+G.K+G.K.$\omega+\alpha$	-7672.70	0.53	0.69		12.45	0.22						0.04*		0.42

Table 5.4: **Statistics from analyses of primate *F8* using the CNF baseline model**. Model – substitution models applied to estimate parameter values; lnL – log-likelihood; $p_1 - p_8$ – probabilities from LRTs, see Figure 5.1 for corresponding null and alternative hypotheses; * – nominally significant at the 0.05 level. See Table 5.1 for definition of terms.

5.3.1 Elevated CpG transition and transversion rate were evident

Incorporation of the dinucleotide effect terms G and $G.K$ significantly improved the fit over the baseline codon substitution model for the majority of primate genes and some yeast genes. LRTs that compared CNF+G versus CNF models and CNF+G.K versus CNF models revealed that 67.81% and 68.02% of primate genes (Table 5.5) had significant support for the addition of the G and $G.K$ terms respectively. These results indicated different mutation properties of CpG-containing codons in primates. In contrast, only 12.20% and 15.87% of yeast genes displayed significant support for the G and $G.K$ terms respectively.

An elevated CpG substitution rate was evident from both example genes and the majority of primate genes. For both the *BRCA1* (Table 5.3) and *F8* (Table 5.4) genes, the $\hat{G}$ value was significantly greater than 1. $\hat{G}$ values (Figure 5.3) from nominally significant CNF+G to CNF models from all primate genes were dominantly and sparsely distributed on the right-hand side of 1, with some values greater than 10. The mode of the G term was around 3 and 80% of $\hat{G}$ values lay between 2 and 6. These results suggested that the CpG substitution rate was higher than the background substitution rate.

The CpG transition rate was also higher than the background transition rate from both example genes and the majority of the primate genes. Examination of the $\hat{G.K}$ value in the CNF+G.K model compared to the CNF model revealed that $\hat{G.K}$ was nominally significant and greater than 1 for both the *BRCA1* (Table 5.3) and *F8* (Table 5.4) genes. A histogram of $\hat{G.K}$ values (Figure 5.4) from genes with statistics supporting a distinct CpG transition rate showed that $\hat{G.K}$ values were mostly greater than 1. The shape of the distribution of $\hat{G.K}$ values was generally similar to that for $\hat{G}$ values with the mode located at approximately 4 and ~70% of the values within the range of 2-6.

Species	Total	Path I		Path II	
		$p_1 < 0.05$	$\arg\max(p_1, p_2) < 0.05$	$p_3 < 0.05$	$\arg\max(p_3, p_4) < 0.05$
Primate	10044	6811 (67.81%)	1900 (27.90%)	6832 (68.02%)	1287 (18.84%)
Yeast	1934	236 (12.20%)	24 (10.17%)	307 (15.87%)	34 (11.07%)

Table 5.5: **Number of significant genes from paths I and II.** The values in the **Total** column represent the number of genes examined under the null hypothesis common to both paths. The percentages are calculated relative to the number of genes in the corresponding null hypothesis

Species	Total	Path III		Path IV	
		$p_5 < 0.05$	$\arg\max(p_5, p_6) < 0.05$	$p_7 < 0.05$	$\arg\max(p_7, p_8) < 0.05$
Primate	6832	1513 (22.41%)	201 (13.13%)	1477 (21.62%)	157 (10.63%)
Yeast	307	143 (46.58%)	12 (8.39%)	63 (20.52%)	24 (38.10%)

Table 5.6: **Number of significant genes from paths III and IV** The values in the **Total** column represent the number of genes examined under the null hypothesis common to both paths. The percentages are calculated relative to the number of genes in the corresponding null hypothesis

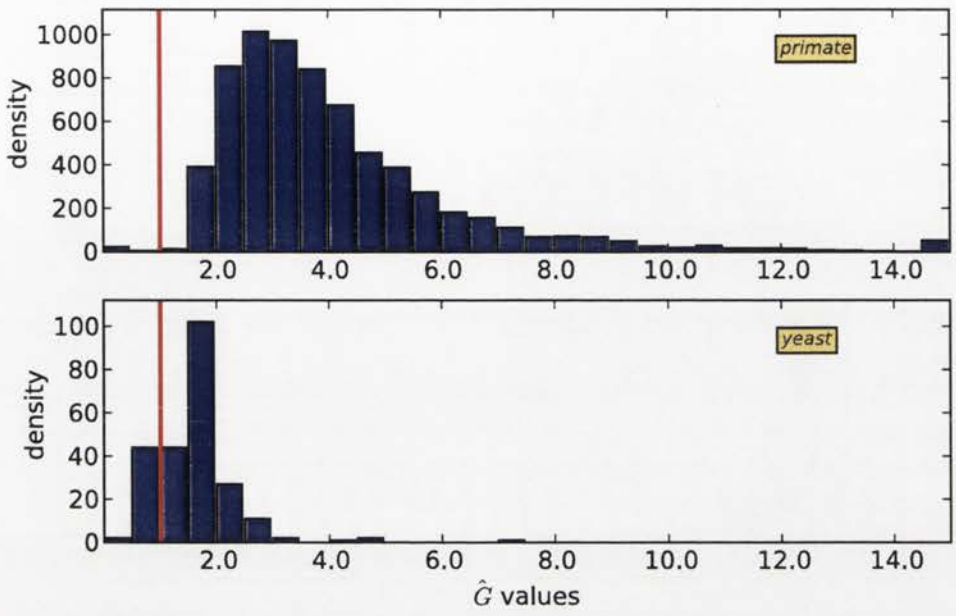


Figure 5.3: **Histogram of estimated $\hat{G}$ distribution from CNF+G model with $p_1 < 0.05$ from primate and yeast genes.** Red vertical lines represent 1 which means there is no effect of the parameter.

An elevated CpG transversion rate was also evident in some primate genes. Since a significant G term from p_1 was not necessarily arising from a different CpG transversion rate to the general transversion rate, addition of the G term to the CNF model does not distinguish the CpG transversion rate from the general transversion rate. Instead, the joint model, CNF+G+G.K, distinguishes CpG transitions and CpG transversions. Identifying the significance of CpG transversion requires comparison of CNF+G+G.K against CNF+G.K. Thus, if p_4 is significant (following significant p_3), the CpG transversion effect is robust, such as for the *BRCA1* gene (Table 5.3), but not the *F8* gene (Table 5.4). We obtained 1287 such genes whose $\hat{G}$ values from the CNF+G+G.K model were predominantly greater than 1 (Figure 5.5).

Yeast genes also exhibited slightly elevated CpG substitution rates. In the case of the G and $G.K$ terms alone, both $\hat{G}$ and $\hat{G.K}$ values were mainly distributed above one within a narrow range (Figure 5.3, 5.4). The mode was around 1.5-2,

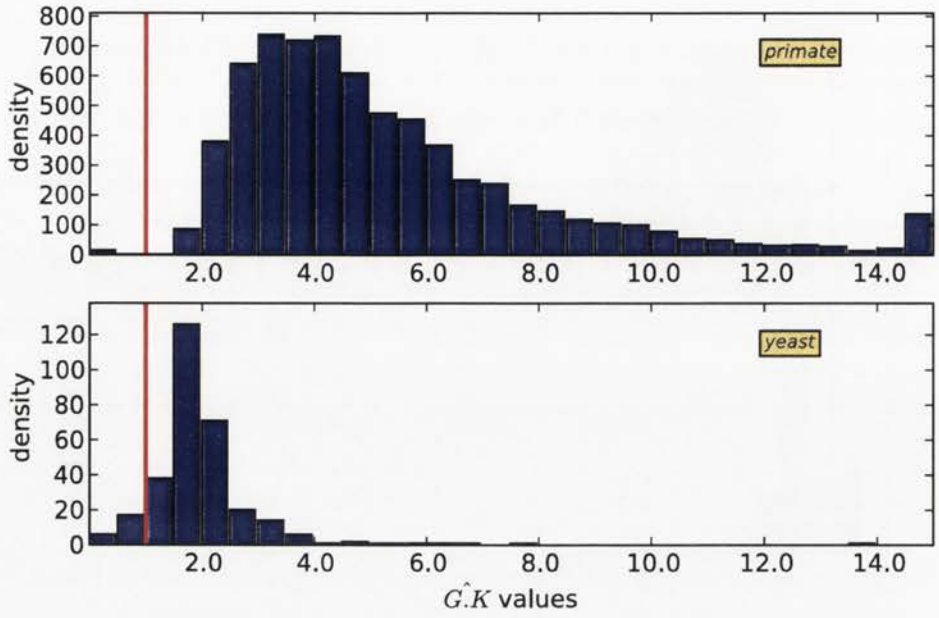


Figure 5.4: **Histogram of estimated $\hat{G.K}$ distribution from CNF+G.K model with $p_3 < 0.05$ from primate and yeast genes.** Red vertical lines represent position 1.

and ~80% of $\hat{G}$ and ~60% $\hat{G.K}$ values were less than 2. For the robustness of CpG transversions, we examined genes with p_3 and p_4 less than 0.05. There were 34 such genes and their $\hat{G}$ values from the CNF+G+G.K model were distributed at both sides of the value 1 (Figure 5.5). These results suggested that only the CpG transition rate was elevated in yeast, but to a much smaller extent than that in primates.

5.3.2 CpG transitions were the major context-dependent effect

Comparison of significant G and $G.K$ values from the path I and path II (Figure 5.1) models revealed that the CpG transition was the major exchange term distinguishing CpG substitutions from background. In path II, improvement of model fitness from CNF+G.K to CNF model was purely from CpG transitions

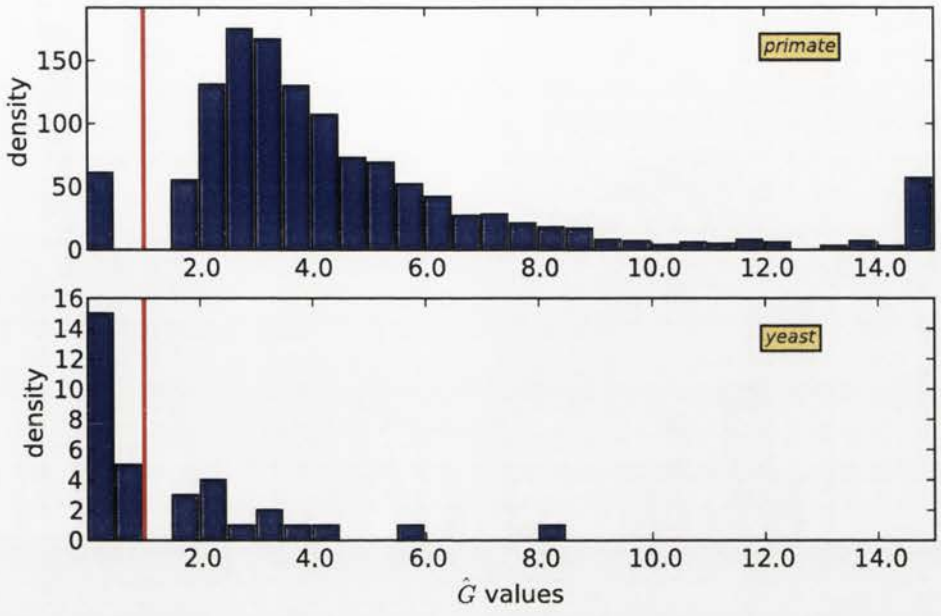


Figure 5.5: **Histogram of estimated $\hat{G}$ distribution from CNF+G+G.K model with** $\arg\max(p_3, p_4) < 0.05$ **from primate and yeast genes.** Red vertical lines represent 1.

defined only by $G.K$. Consequently, if p_3 was significant but p_2 was not, we still considered that the CpG transition estimate was robust, such as for the $F8$ gene. In path I, since $G.K$ is a subset of the G term, CpG transitions contributed to the increased likelihood for the CNF+G compared to the CNF model. The additional $G.K$ term added to the CNF+G model distinguished those with exceptional CpG transition effects. This was evident from the $\hat{G.K}$ values, since there were a large number of genes with $\hat{G.K}$ values greater than 10 (Figure 5.6), from the CNF+G+G.K model with nominally significant p_1 and p_2. Thus, for genes displaying a differential CpG substitution rate from the background, the major effect was CpG transitions in both primates and yeast.

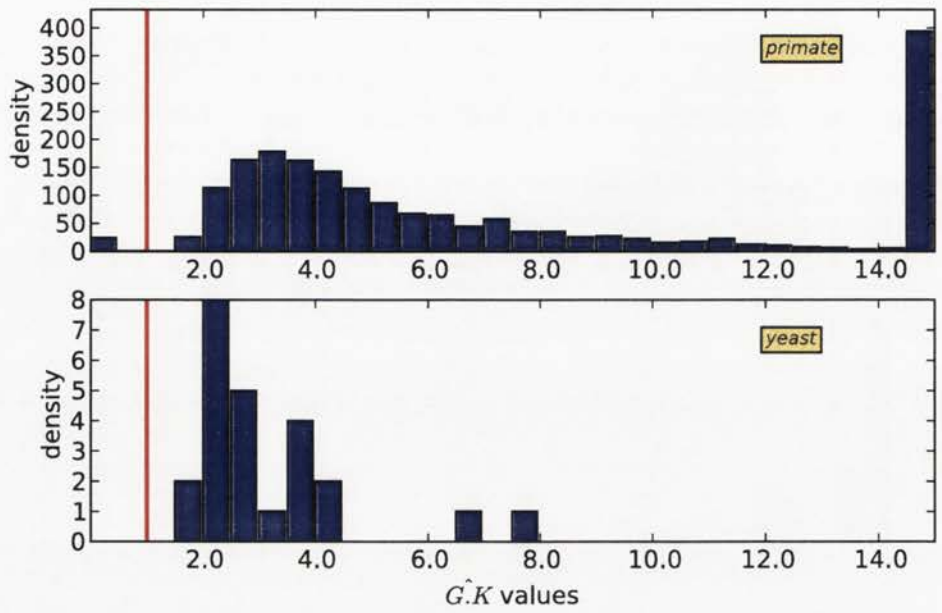


Figure 5.6: **Histogram of estimated $\hat{G.K}$ distribution from CNF+G+G.K model with $\arg\max(p_1, p_2) < 0.05$ from primate and yeast genes.** Red vertical lines represent 1.

5.3.3 Methylation-affected amino acids exhibited a different nonsynonymous substitution rate

That $\hat{\alpha}$ values were generally less than 1 in both primates and yeast suggested less permissive amino acid exchanges in MAA than that in non-MAA. We first assessed whether MAAs undergo distinctive selective strength from other amino acids as measured by the parameter α using LRTs comparing CNF+G.K+α versus CNF+G.K (Table 5.6). We obtained 1,513 primate and 143 yeast genes with nominally significant support for α, whose values were mostly less than 1 (Figure 5.7). Since α includes both CpG and non-CpG events, we further examined the $\hat{\alpha}$ values from genes nominally significant for p_3, p_7 and p_8, which were mainly contributed by non-CpG events in MAAs. We also obtained similar $\hat{\alpha}$ distributions (Figure 5.8). The relatively small numbers of genes from this step were likely due to reduced statistical power from multiple tests. Thus, amino

acid exchanges defined by MAAs had a different nonsynonymous substitution rate to other amino acids potentially arising from their distinct physico-chemical properties.

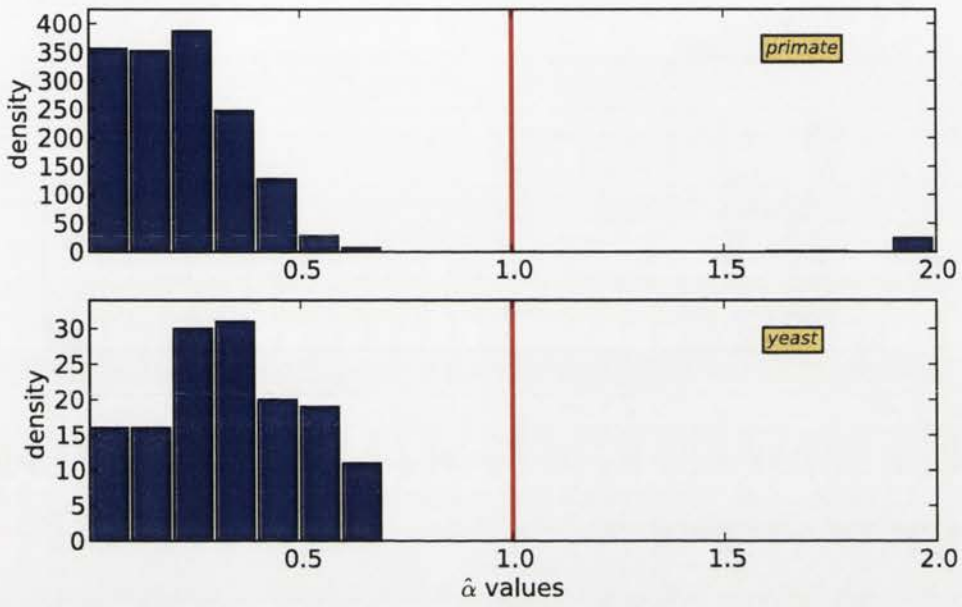


Figure 5.7: **Histogram of estimated $\hat{\alpha}$ distribution from CNF+G.K+α model with $\arg\max(p_3, p_5) < 0.05$ from primate and yeast genes.** Red vertical lines represent 1.

5.3.4 CpG-encoded amino acids were subjected to stronger purifying selection in primates than in yeast

CpG-encoded amino acids exchanges corresponding with those affected by $G.K.\omega$ exhibited significantly distinct different strengths of natural selection from other amino acid exchanges, but the modes were different between primates and yeast. We considered the statistical support for $G.K.\omega$ from a LRT by first comparing CNF+G.K+G.K.ω versus CNF+G.K because its effect arises purely from non-synonymous CpG transitions. We obtained 1,477 primate and 63 yeast genes with nominally significant p_3 and p_7 (Table 5.6), in which $G.\hat{K}.\omega$ values (Fig-

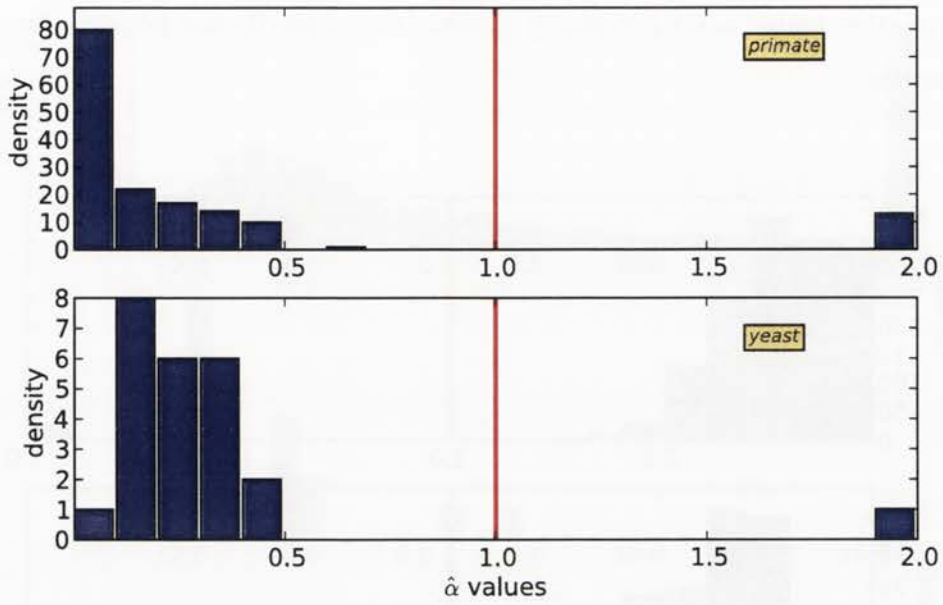


Figure 5.8: **Histogram of estimated $\hat{\alpha}$ distribution from CNF+G.K+G.K.ω + α model with** $\arg\max(p_3, p_7, p_8) < 0.05$ **from primate and yeast genes.** Red vertical lines represent 1.

ure 5.10) were predominantly less than 1. However, at this stage, whether the stronger purifying selection exhibited by $G.K.\omega$ is due to elevated CpG mutability or generic amino acid physico-chemical properties is unknown. Therefore, we further examined $G.K.\omega$ with background selective strength measured by α. This will distinguish CpG events from non-CpG events among MAA exchanges. The LRT for this purpose was CNF+G.K+α+G.K.ω versus CNF+G.K+α. We obtained 201 primate and 12 yeast significant genes in this step, of which 13 primate genes, but no yeast genes, remained significant after multiple test correction [86]. For genes exhibiting nominally significant p_3, p_5, and p_6, $G.\hat{K}.\omega$ (Figure 5.9) was still predominantly less than 1 in primates, but distributed on both sides of 1 in yeast. These results were consistent with the CpG-encoded amino acids having experienced stronger purifying selection in primates, but not in yeast.

The *BRCA1* and *F8* genes (Table 5.3 and 5.4) displayed different selective constraints on CpG-containing codons. For the *BRCA1* gene, both α and $G.K.\omega$

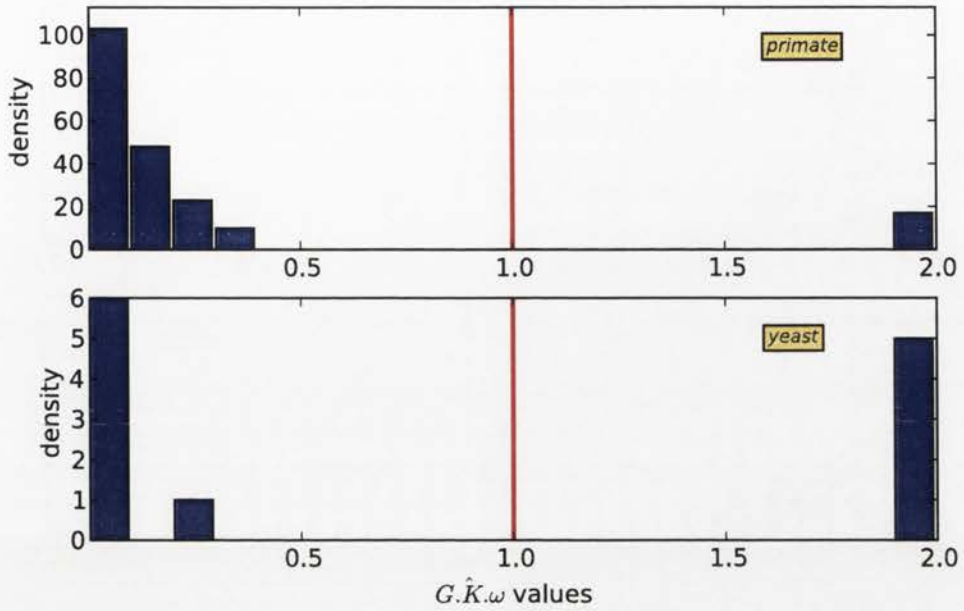


Figure 5.9: **Histogram of estimated $G.\hat{K}.\omega$ distribution from CNF+G.K+G.K.ω+α model with** $\arg\max(p_3, p_5, p_6) < 0.05$ **from primate and yeast genes.** red vertical lines represent 1.

were not significantly supported from paths III and IV, although the CpG transition rate was greatly elevated. For the *F8* gene, we obtained significant support for α from p_5 and significant support for $G.K.\omega$ from p_6 and p_7. These results suggested that CpG-encoded amino acids in the *F8* gene occupy important functional positions, which was consistent with clinical observations that mutations in CpG-containing codons were considered highly likely to cause disease [147].

5.3.5 Genes displaying significant CpG effect were enriched in disease-causing genes

Among sampled human genes that have OMIM records, genes with evidence for strong purifying selection opposing nonsynonymous CpG transitions were significantly enriched in OMIM allelic genes. The criteria used to identify the genes with strong purifying selection opposing CpG transitions were: p_3 and $p_7 < 0.05$,

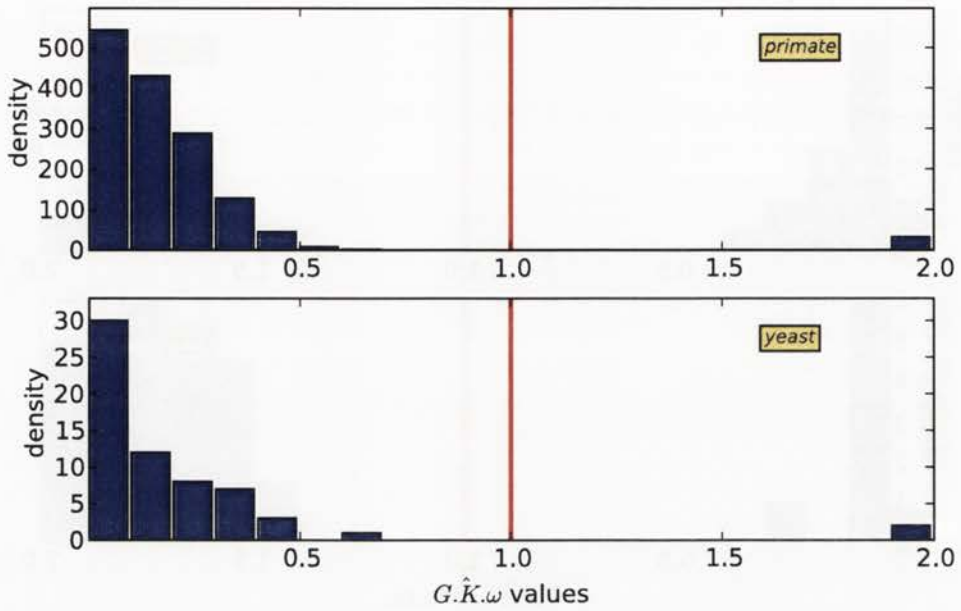


Figure 5.10: **Histogram of estimated $G.\hat{K}.\omega$ distribution from CNF+G.K+G.K.ω model with** $\arg\max(p_3, p_7) < 0.05$ **from primate and yeast genes.** red vertical lines represent 1.

$G.K > 1$, and $G.K.\omega < 1$. Thus, if nonsynonymous mutations occur within a CpG context, they have a higher chance to negatively impair phenotype, and thus to cause disease. Among sampled OMIM genes, 986 genes displayed significant support for differential $G.K$ and $G.K.\omega$ with 241 genes being classified as disease-associated. Accordingly, the sampled genes were classified into four categories (Table 5.7). Using a Fisher's exact test, the genes whose CpG codons were under stronger purifying natural selection were significantly enriched for association with disease. These results are consistent with deleterious phenotypic consequences arising from CpG mutations.

Table 5.7: **Testing of enrichment of genes displaying significant CpG effect in OMIM disease-associated genes by Fisher's exact test**

	Non-disease associated	Disease associated	p
Non-CpG effect	4520	1193	0.0072
CpG effect	745	241	

5.3.6 Within-species genetic variation analyses further supported purifying selection affecting exonic CpG polymorphism

The numbers of CpG-associated SNPs are shown in Table 5.8. To make SNP numbers comparable between each category, the number of SNPs was normalized by the number of codon pairs involving one nucleotide substitution for each class, denoted as SNPs per pair (Table 5.8). For example, for synonymous transition SNPs of CpG-containing codons, there are four exchanges TCG $\leftrightarrow$ TCA (Ser), CCG $\leftrightarrow$ CCA (Pro), ACG $\leftrightarrow$ ACA (Thr), and GCG $\leftrightarrow$ GCA (Ala). As a result, the SNPs per pair under this class were 6,621 (by dividing 26,484 by 4).

SNPs within a CpG context comprised a significant proportion of all alleles in coding sequences. Overall, $\sim$40% of codon SNPs were associated with CpG sites, and $\sim$35.7% of nonsynonymous variants occurred within a CpG context. This is substantially reduced relative to the $\sim$45.6% of synonymous SNPs within a CpG context.

Transition exchange alleles were the major component distinguishing CpG SNPs from other dinucleotide contexts. Elevated transitions in CpG-containing SNPs were evident for both synonymous and nonsynonymous alleles. The number of SNPs per pair for CpG transition alleles was $\sim$ 6 (6865/1170) fold and $\sim$ 3.5 (2500/727) fold that of non-CpG alleles at synonymous and nonsynonymous sites respectively. Transversion exchange alleles were also higher in CpG SNPs, but

were much less significant than transitions. The number of CpG transversion alleles per pair was $\sim$ 1.8 fold higher on synonymous and $\sim$ 1.6 fold higher on nonsynonymous than for non-CpG transversion alleles.

Natural selection also appeared to have a stronger effect on CpG nonsynonymous substitutions. Assuming the mutation rate is the same for CpG substitutions at synonymous and nonsynonymous sites, the transition to transversion ratio (λ) should be the same for nonsynonymous and synonymous CpG SNPs. However, λ at nonsynonymous CpG SNPs is 0.49 (5.98/12.09) of the value of λ at synonymous CpG SNPs. The lower λ value of nonsynonymous CpG SNPs is consistent with the stronger purifying influence of natural selection. This result suggested substantially reduced fitness for CpG transitions at nonsynonymous sites.

5.4 Discussion

Our analyses firmly support a 5^mC-derived shift in mutation-selection balance on CpG in primates and elucidated the consequences of such changes on phenotype. Consistent with methylation-mediated mutations, the CpG transition rate was accelerated for a majority of primate genes. An elevated CpG transversion rate was also evident from a small number of primate genes, suggesting that other factors are involved. Along with greater mutation pressure, selection constraints on CpG-encoded amino acids were significantly stronger than in the same set of amino acids encoded by non-CpG codons, with the main effect being purifying selection. Furthermore, genes with an elevated CpG transition rate but stronger purifying selection on CpG codons were enriched in disease-causing genes. These observations support our hypothesis that CpG codons are more functionally important and phenotypically influential. In contrast to primates, CpG codons in yeast exhibited a much weaker mutation pressure, and there was no evidence showing stronger purifying selection specifically operating on CpG codons.

Table 5.8: **Classification of human biallelic SNPs within coding regions**

Effect type	**CpG context**	**substitution type**	**Codon pairs**	**SNP number**	**SNPs per pair**	λ
Nonsynonymous	CpG	transition	11	27497	2500	5.98
		transversion	20	8354	418	
	Non-CpG	transition	47	34172	727	2.82
		transversion	118	30430	258	
Synonymous	CpG	transition	4	27460	6865	12.09
		transversion	10	5680	568	
	Non-CpG	transition	27	31588	1170	3.80
		transversion	26	7999	308	

A substantially higher CpG transition rate than background transition rate for the majority of primate genes confirmed that transitions are the major methylation-induced mutations. Since spontaneous deamination of 5^mC results in T, CpG transitions can simply arise from un-repaired T/G mismatches during DNA replication or repair on the opposite strand, which produces a T/A pair from the original C/G pair. Both situations create permanent transition mutations in the cell. Thus, the observed dominant transition effect on CpGs is concordant with the expected methylation-derived mutations.

An elevated CpG transversion rate in a small number of primate genes indicates additional causes other than 5^mC deamination. One possible cause is base misincorporation during DNA repair. Compared with DNA replication, DNA repair processes tend to be error-prone, which is possibly due to the use of low-fidelity DNA polymerases [148]. Thus, the high 5^mC deamination rate may be accompanied with a high probability of complete replacement during DNA repair and lead to a high CpG transversion rate. The other possibility is that CpG sites are DNA damage hotspots irrespective of methylation status. For instance, CpG has been identified as a preferred target of oxidative damage [149]. Since oxidative reactions are one of the major mechanisms of DNA damage and some of these predominantly produce transversions, e.g.8-OH-dG [72], this preference may cause an elevated CpG transversion rate.

Stronger purifying selection operating on CpG-encoded amino acids suggested that these amino acids are more likely to affect trait evolution. That α was predominantly less than 1 for both primates and yeast reveals that the amino acid exchanges involved in MAA are generally less permissive than other amino acids, and that it was necessary to have α in the model. With appropriately adjusted background selective constraints modeled by α, CpG codons were further distinguished from non-CpG codons by stronger purifying selection. A further decrease in permissiveness on CpG nonsynonymous exchanges indicates that these amino acids occupy functionally significant positions. This conjecture was fully

supported by enrichment of disease-causing genes in CpG-effected genes. Thus, our analyses provide a new perspective in identifying codons that potentially affect phenotype.

Different CpG mutation and selection properties between primates and yeast are apparent, generally consistent with the expected outcomes from methylated and non-methylated genomes. If there is no impact from other evolutionary forces, CpG sites should evolve at the same rate as that of the background for genomes free of methylation. Thus, we expected a small number of yeast genes displaying statistical support for G, $G.K$ and $G.K.\omega$ terms with values equally distributed on both sides of 1. Concordantly, only a small proportion of yeast genes, (which was not due to a lack of statistical power (supplementary and Figure 5.11)), showed significant support for differential $G.K$ and G with their values distributed much lower than those from primates. For the measure of selection constraints on CpG codons in yeast, the number of genes with nominally significant $G.K.\omega$ after adjusting by α was extremely low and distributed on both sides of 1. These observations clearly revealed distinct mutation-selection balance between primates and yeast due to different methylation status. However, yeast genes also displayed weak mutation pressure on CpG codons given that $\hat{G.K}$ was primarily greater than 1. This result suggested that other evolutionary forces are acting on yeast CpG codons.

The causes of the slightly elevated CpG transition rate in yeast are open to question. One potential cause is the codon usage bias in yeast. According to Bennetzen and Hall (1982) [150], for four-fold degenerate sites like XY(U/C/A/G), XYA and XYG codons are rarely used. Three out of five amino acids encoded by CpG-containing codons, namely Serine(TCG), Threonine(ACG), and Alanine(GCG), belong to this category. Another observation was that although overall CpG is normally represented in *S. cerevisiae* [132], CpG is substantially suppressed at codon positions I-II and positions II-III [151]. Similar analyses by Kliman et. al (2003) [152] also revealed that none of the CpG-containing codons were the

major synonymous codons in yeast, while CpA tended to comprise the major synonymous codons. Thus the less-preferred CpG codons may undergo mutation pressure to preferred synonymous codons such as CpA, which may result in a higher transition rate at CpG sites. The other cause may also be preferential DNA damage at CpG as mentioned above. Another intuitive explanation is also related to cytosine methylation. Some yeast, e.g. *Kluyveromyces lactis* and *Candida albicans*, have well-established methylation systems and exhibit CpG suppression in their genomes [132]. Although brewers yeast [146] is generally regarded as free of methylation, it may have low levels of cytosine methylation that cannot be easily detected, or may undergo a methylated state in certain developmental stages [153], or have experienced a methylation stage in its evolutionary history. Overall, the exact cause of the elevated CpG transition rate in yeast needs further examination.

Use of the CNF model as baseline corrected the systematic bias from previous analyses [135] and led to a more sensible interpretation. The choice of baseline codon substitution models has a great impact on parameter estimation. Using the Y98 model as a baseline model, $\hat{G}$ became considerably higher while $\hat{G.K}$ became lower than those estimates from a baseline CNF model for the *BRCA1* gene (see supplementary). With an extremely high value of $\hat{G}$ and a relatively small $\hat{G.K}$, CpG transversion seems an indispensable factor in defining CpG mutation properties. Consequently, Huttley used $G.\omega$ (the interaction of the G term and the ω term) to measure selective strength on CpG codons, while according to modeling based on the CNF model, a $G.K.\omega$ term was more appropriate. Thus, depending on sequence compositions, a biased Y98 model produces confounded parameter estimates when compared with the CNF model and may lead to misinterpretation.

Although the codon substitution model approach is superior to most other methods, it is imperfect because CpG mutations from neighboring codons are not considered. The codon substitution model applied here assumes that each codon

column in the alignment evolves independently. Thus, CpG dinucleotides that span codon boundaries (NNC GNN) are not considered. This will affect the estimates of CpG effect, especially for primates since the evolutionary rate at CpG sites clearly differs from that of the background. Notably, the SNP counting exercise was not affected by codon boundaries, but it ignores substitution rate and selection heterogeneity among genes, and lacks formal statistical tests. These defects make the codon substitution model a better choice. An alternative approach may be the combination of a codon substitution model and a HMM to model the dependence of neighboring states. However, the question of how to model transition probabilities on neighboring codons within a CpG and non-CpG context needs further examination.

The number of genes that showed a significant CpG effect was limited by statistical power in primates. Since the sampled yeast genes diverge further than primate genes, the statistical power for yeast is much stronger than that for primates (supplementary). It was clear that when the CpG transition rate is 2-fold higher than the background, it has a 98% chance of being detected in yeast but only an ~60% chance in primates (Figure 5.11). In addition, sequential LRTs further reduce the number of significant genes at later steps. A compensation for the lack of statistical power in primates is the alignment length. With more columns to be counted, larger genes are more likely to be significant in LRTs. This was evident in that CpG-affected genes with (or without) disease-association were generally larger than overall OMIM allelic genes (data not shown). This limitation will result in a failure to detect some smaller genes with a significant CpG effect.

5.5 Supplementary

5.5.1 Modeling based on the Y98 model

The same analysis procedure was followed using the Y98 model instead of the CNF model as the baseline. The Y98 model [54] is a modified GY94 model [133] which uses codon equilibrium frequencies, a parameter λ to represent background nucleotide substitution process, and a parameter ω to represent selection pressure on nonsynonymous codons. According to Yap et al [136], this parameterization causes less bias than other modified GY94 models, which makes it comparable to the CNF model. Following the conventional definition, q_{ij}, the relative exchange rate from codon i to codon j in Y98, is defined as:

$$q_{ij,i\neq j} = \begin{cases} 0 & \text{more than one nucleotide difference} \\ \pi_j & \text{synonymous transversion} \\ \pi_j \cdot \lambda & \text{synonymous transition} \\ \pi_j \cdot \omega & \text{nonsynonymous transversion} \\ \pi_j \cdot \lambda \cdot \omega & \text{nonsynonymous transition} \end{cases} \tag{5.8}$$

where π_j is the equilibrium frequency of codon j. The addition of other parameters (G, $G.K$, α, and $G.K.\omega$) was the same as using the baseline CNF model. The resulting parameter estimates for the *BRCA1* and *F8* genes are displayed in Table 5.9 and 5.10 respectively.

Model	**lnL**	$\hat{\omega}$	$\hat{\alpha}$	$\hat{G}$	$\hat{G.K}$	$\hat{G.K.\omega}$	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
Y98	-8833.33	0.58												
Y98+G	-8783.68	0.58		12.33			0.00*							
Y98+G.K	-8793.08	0.55			13.84				0.00*					
Y98+G+G.K	-8782.86	0.57		8.62	1.66			0.20		0.00*				
Y98+G.K+α	-8792.98	0.56	0.87		15.06						0.65			
Y98+G.K+G.K.ω	-8792.96	0.56			15.97	0.80							0.62	
Y98+G.K+G.K.$\omega + \alpha$	-8792.94	0.56	0.93		15.96	0.86						0.79		0.85

Table 5.9: **Statistics from analyses of primate *BRCA1* using the Y98 baseline model**. Model – substitution models applied to estimate parameter values; lnL – log-likelihood; $p_1 - p_8$ – probabilities from corresponding LRTs; * – nominally significant at the 0.05 level. See Table 5.1 for definition of terms.

Model	**lnL**	$\hat{\omega}$	$\hat{\alpha}$	$\hat{G}$	$G.\hat{K}$	$G.\hat{K}.\omega$	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
Y98	-7685.89	0.47												
Y98+G	-7672.69	0.44		5.12			0.00*							
Y98+G.K	-7674.35	0.43			6.33				0.00*					
Y98+G+G.K	-7671.93	0.43		3.40	1.89			0.22		0.03*				
Y98+G.K+α	-7672.23	0.48	0.42		10.76						0.04*			
Y98+G.K+G.K.ω	-7669.68	0.50			16.95	0.18							0.00*	
Y98+G.K+G.K.$\omega + \alpha$	-7669.67	0.50	0.95		16.94	0.19						0.02*		0.91

Table 5.10: **Statistics from analyses of primate *F8* using the Y98 baseline model**. Model – substitution models applied to estimate parameter values; lnL – log-likelihood; $p_1 - p_8$ – probabilities from corresponding LRTs; * – nominally significant at the 0.05 level. See Table 5.1 for definition of terms.

5.5.2 Assessment of statistical power

Simulations were based on the CNF model with parameter values estimated from sampled alignments and additional fixed $G.K$ values. 100 alignments were randomly selected from primates and yeast respectively. Each alignment was fitted to a CNF model. The resulting parameter values (including GTR, ω, branch lengths, and codon frequencies) and $G.K$ (whose values range from 1.5 to 3.0) were used to simulate alignments. Simulated alignment lengths were equal to the observed sampled alignment lengths. Under each $\hat{G.K}$, 200 alignments were simulated and a LRT was performed to compare CNF+G.K versus CNF. Statistical power was measured as the percentage of detected significant $G.K$ effects under each condition. Simulations revealed that the LRT was able to correctly identify an elevated CpG transition rate in primates and yeast. Figure 5.11 reports the statistical power of the LRT for the $G.K$ effect with different $\hat{G.K}$ values. The statistical power was generally greater at each $\hat{G.K}$ from yeast alignments than from primate alignments. For yeast, a LRT was able to detect when $\hat{G.K}$ was equal to 2.0 more than 95% of the time, while the same power was not reached until $\hat{G.K}$ equal to 3.0 in primates. Thus, the relatively small number of significant genes from path I and path II in yeast was not due to a lack of statistical power.

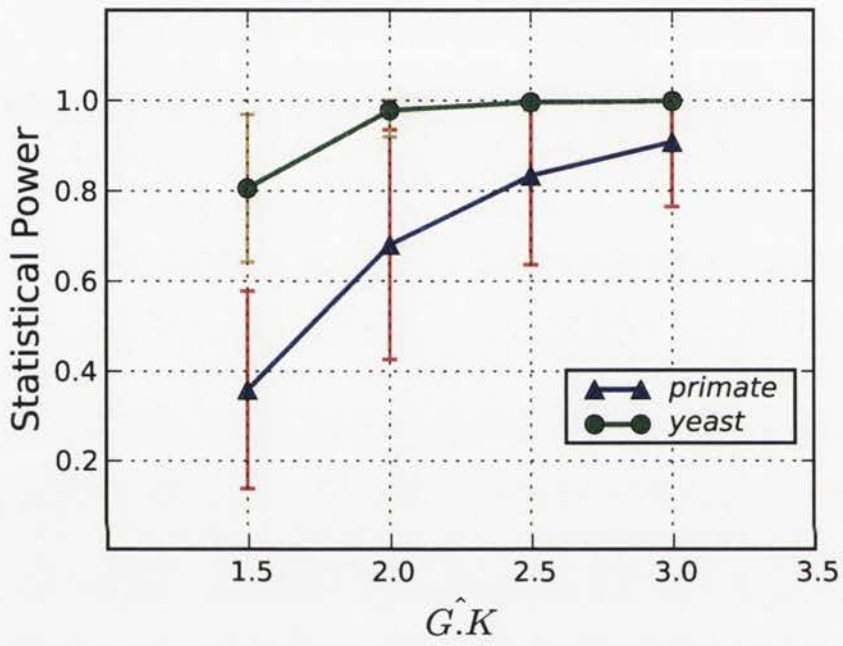


Figure 5.11: **Statistical power of LRTs from simulated primate and yeast alignments.** LRTs compared CNF+G.K model versus CNF model at different $G\hat{.}K$. 100 alignments were randomly selected to estimate parameter values of CNF model from each taxa and each alignment were simulated 200 times. The blue triangles and green circles represent the mean statistical power at each $G\hat{.}K$ for primates and yeast respectively. The red and yellow vertical bars represent corresponding standard deviation.

Conclusion

The main focus of the research reported in this thesis has been the influence of epigenetic factors on sequence evolution. The two specifically selected epigenetic factors, namely chromatin structure and DNA methylation, affect mutagenesis in unique ways that are studied in this thesis. It was found that DHSs with their open chromatin structure exhibited a reduced substitution rate and different substitution profiles compared to relatively closed Flank regions, and nucleosome placement induced localised substitution rate heterogeneity. For 5^mC, our analyses systematically evaluated its increased mutational potential in both transitions and transversions, and the stronger purifying selection on CpG-encoded amino acids derived from the elevated 5^mC transition rate. These results are in line with recent reports that considered epigenetic factors to be important contributors to substitution rate heterogeneity across genomes.

While a nucleosome code has been proposed, the periodic pattern in substitution spectra putatively arising from nucleosome positioning has important implications for identifying nucleosome positions. It has been suggested that DNA encodes the positions of nucleosomes with a 10 bp periodicity of AA/TT/TA [154]. This sequence feature was thought to be required for sharp bending of DNA on the surface of histones. However, whether this rule can be applied universally to individual nucleosomes is still under investigation ([155], Epps, Ying, and Huttley unpublished data). In contrast, the nucleosome footprint represented by an ~200 bp periodicity from sequence comparisons was derived from local genomic regions.

Additionally, the larger than expected peak width from substitution spectra may represent fuzzy nucleosomes or activities such as chromatin remodeling. Thus, it may be possible to identify nucleosome positioning and activities from comparative genomic analyses when further genome sequences become available and more accurate nucleosome mapping data can be used for comparisons.

It would be interesting to know whether there are any forms of selection operating on nucleosome positioning signals. Our results and other reports [11, 13, 14, 15] all suggest that nucleosomal sequences evolve faster than linker sites. Given that nucleosome positions are essential in regulating gene expression and other important nuclear activities, whether mutations within a stretch of nucleosome associated-DNA will change the sequence affinity for nucleosome binding and invoke natural selection is unknown. Moreover, similar to the redundancy in the genetic code in which 61 codons encode 20 amino acids, the nucleosome positioning code is very likely to be degenerate. Besides the sequence, such a code may exist on the secondary DNA structure which leads to great plasticity as long as the histone binding preference can be preserved. Therefore, if natural selection does operate on nucleosomes, it is of particular interest to understand whether its dominant effect is on the primary DNA sequence, or DNA conformation.

For the second of the epigenetic factors examined, DNA methylation, my results suggested that it is practical to identify genes with CpG codons occupying critical functional sites. Such a possibility was indicated by stronger purifying selection on CpG-encoded amino acids than other amino acids and further strengthened by the observation of enrichment of disease-association in CpG-affected genes. In these analyses, closely related species were preferred since they are likely to share similar DNA repair systems and protein functions. However, this preference greatly reduced the statistical power from sequential LRTs. One solution is to include more species to compensate for the loss of power. With more primate genomes being sequenced, we expect to obtain more accurate estimates for these analyses and detect greater numbers of genes displaying a significant CpG

effect.

Other than whole genes, it would be particularly useful to identify CpG-encoded amino acids that affect phenotype. Since amino acids occupying different positions within a protein undergo different selective pressures, selective constraints on CpG codons will be heterogenous for a gene. With $\sim$ 40% of coding SNPs located in a CpG context, the problem of how to distinguish CpG mutations that cause phenotypical changes, such as disease, from neutral SNPs is an interesting problem to be tackled. One possible solution is to apply a Phylo-HMM to detect conserved CpG codons within highly conserved protein domains.

In conclusion, the outcomes presented in this thesis significantly improve our understanding of the effect of epigenetic factors on substitution. They shed light on the correlation between sequence evolution and epigenetic states. With the continued development of comparative genomics algorithms and the availability of more sequencing data, it is becoming practical to predict physical features and draw functional inferences readily from epigenetic states.

Bibliography

[1] Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T: **Male-driven molecular evolution: a model and nucleotide sequence analysis.** *Cold Spring Harb Symp Quant Biol* 1987, **52**:863–867.

[2] Shimmin LC, Chang BH, Li WH: **Male-driven evolution of DNA sequences.** *Nature* 1993, **362**:745–747.

[3] Lercher MJ, Williams EJ, Hurst LD: **Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias.** *Mol Biol Evol* 2001, **18**:2032–2039.

[4] Arndt PF, Hwa T, Petrov DA: **Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects.** *J Mol Evol* 2005, **60**:748–763.

[5] Gaffney DJ, Keightley PD: **The scale of mutational variation in the murid genome.** *Genome Res* 2005, **15**:1086–1094.

[6] Piganeau G, Mouchiroud D, Duret L, Gautier C: **Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores.** *J Mol Evol* 2002, **54**:129–133.

[7] Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature* 2001, **409**:928–933.

[8] Prendergast JGD, Campbell H, Gilbert N, Dunlop MG, Bickmore WA, Semple CAM: **Chromatin structure and evolution in the human genome.** *BMC Evol Biol* 2007, **7**:72.

[9] Li WH: *Molecular Evolution.* Sunderland, Mass.: Sinauer Associates 1997.

[10] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15**:1034–1050.

[11] Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ: **Genome-scale identification of nucleosome positions in S. cerevisiae.** *Science* 2005, **309**:626–630.

[12] Ozsolak F, Song JS, Liu XS, Fisher DE: **High-throughput mapping of the chromatin structure of human promoters.** *Nat Biotechnol* 2007, **25**:244–248.

[13] Washietl S, Machne R, Goldman N: **Evolutionary footprints of nucleosome positions in yeast.** *Trends Genet* 2008, **24**:583–587.

[14] Warnecke T, Batada NN, Hurst LD: **The impact of the nucleosome code on protein-coding sequence evolution in yeast.** *PLoS Genet* 2008, **4**:e1000250.

[15] Sasaki S, Mello CC, Shimada A, Nakatani Y, Hashimoto SI, Ogawa M, Matsushima K, Gu SG, Kasahara M, Ahsan B, Sasaki A, Saito T, Suzuki Y, Sugano S, Kohara Y, Takeda H, Fire A, Morishita S: **Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites.** *Science* 2009, **323**:401–404.

[16] Sinha NK, Haimes MD: **Molecular mechanisms of substitution mutagenesis. An experimental test of the Watson-Crick and topal-fresco models of base mispairings.** *J Biol Chem* 1981, **256**:10671–10683.

[17] Gojobori T, Li WH, Graur D: **Patterns of nucleotide substitution in pseudogenes and functional genes.** *J Mol Evol* 1982, **18**:360–369.

[18] Coulondre C, Miller JH, Farabaugh PJ, Gilbert W: **Molecular basis of base substitution hotspots in Escherichia coli.** *Nature* 1978, **274**:775–780.

[19] Cooper DN, Youssoufian H: **The CpG dinucleotide and human genetic disease.** *Hum Genet* 1988, **78**:151–155.

[20] Giannelli F, Anagnostopoulos T, Green PM: **Mutation rates in humans. II. Sporadic mutation-specific rates and rate of detrimental human mutations inferred from hemophilia B.** *Am J Hum Genet* 1999, **65**:1580–1587.

[21] Nachman MW, Crowell SL: **Estimate of the mutation rate per nucleotide in humans.** *Genetics* 2000, **156**:297–304.

[22] WATSON JD, CRICK FH: **Genetical implications of the structure of deoxyribonucleic acid.** *Nature* 1953, **171**:964–967.

[23] Topal MD, Fresco JR: **Complementary base pairing and the origin of substitution mutations.** *Nature* 1976, **263**:285–289.

[24] Rosenberg MS, Subramanian S, Kumar S: **Patterns of transitional mutation biases within and among mammalian genomes.** *Mol Biol Evol* 2003, **20**:988–993.

[25] Cadet J, Anselmino C, Douki T, Voituriez L: **Photochemistry of nucleic acids in cells.** *J Photochem Photobiol B* 1992, **15**:277–298.

[26] Tornaletti S, Pfeifer GP: **UV damage and repair mechanisms in mammalian cells.** *Bioessays* 1996, **18**:221–228.

[27] Carty MP, Hauser J, Levine AS, Dixon K: **Replication and mutagenesis of UV-damaged DNA templates in human and monkey cell extracts.** *Mol Cell Biol* 1993, **13**:533–542.

[28] Carty MP, el Saleh S, Zernik-Kobak M, Dixon K: **Analysis of mutations induced by replication of UV-damaged plasmid DNA in HeLa cell extracts.** *Environ Mol Mutagen* 1995, **26**:139–146.

[29] Iyer RR, Pluciennik A, Burdett V, Modrich PL: **DNA mismatch repair: functions and mechanisms.** *Chem Rev* 2006, **106**:302–323.

[30] Petranovic M, Vlahovic K, Zahradka D, Dzidic S, Radman M: **Mismatch repair in xenopus egg extracts is not strand-directed by DNA methylation.** *Neoplasma* 2000, **47**:375–381.

[31] Sancar A: **DNA excision repair.** *Annu Rev Biochem* 1996, **65**:43–81.

[32] Wood RD: **Nucleotide excision repair in mammalian cells.** *J Biol Chem* 1997, **272**:23465–23468.

[33] Kow YW: **Repair of deaminated bases in DNA.** *Free Radic Biol Med* 2002, **33**:886–893.

[34] Mellon I, Spivak G, Hanawalt PC: **Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian DHFR gene.** *Cell* 1987, **51**:241–249.

[35] Feng Z, Hu W, Komissarova E, Pao A, Hung MC, Adair GM, Tang Ms: **Transcription-coupled DNA repair is genomic context-dependent.** *J Biol Chem* 2002, **277**:12777–12783.

[36] Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM: **Phylogenetic shadowing of primate sequences to find functional regions of the human genome.** *Science* 2003, **299**:1391–1394.

[37] Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science* 2004, **304**:1321–1325.

[38] Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJK, Cooke JE, Elgar G: **Highly conserved non-coding sequences are associated with vertebrate development.** *PLoS Biol* 2005, **3**:e7.

[39] Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage

D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**:520–562.

[40] Drummond JT, Bellacosa A: **Human DNA mismatch repair in vitro operates independently of methylation status at CpG sites.** *Nucleic Acids Res* 2001, **29**:2234–2243.

[41] Ng HH, Bird A: **DNA methylation and chromatin modification.** *Curr Opin Genet Dev* 1999, **9**:158–163.

[42] Leonhardt H, Cardoso MC: **DNA methylation, nuclear structure, gene expression and cancer.** *J Cell Biochem Suppl* 2000, **Suppl 35**:78–83.

[43] Robertson KD: **DNA methylation and chromatin - unraveling the tangled web.** *Oncogene* 2002, **21**:5361–5379.

[44] Adams RL, Davis T, Rinaldi A, Eason R: **CpG deficiency, dinucleotide distributions and nucleosome positioning.** *Eur J Biochem* 1987, **165**:107–115.

[45] Mathews CK: **DNA precursor metabolism and genomic stability.** *FASEB J* 2006, **20**:1300–1314.

[46] Bebenek K, Roberts JD, Kunkel TA: **The effects of dNTP pool imbalances on frameshift fidelity during DNA replication.** *J Biol Chem* 1992, **267**:3589–3596.

[47] Wolfe KH, Sharp PM, Li WH: **Mutation rates differ among regions of the mammalian genome.** *Nature* 1989, **337**:283–285.

[48] Huttley GA, Jakobsen IB, Wilson SR, Easteal S: **How important is DNA replication for mutagenesis?** *Mol Biol Evol* 2000, **17**:929–937.

[49] Driscoll DJ, Migeon BR: **Sex difference in methylation of single-copy genes in human meiotic germ cells: implications for X chromosome inactivation, parental imprinting, and origin of CpG mutations.** *Somat Cell Mol Genet* 1990, **16**:267–282.

[50] Kimura M: *The Neutral Theory of Molecular Evolution.* Cambridge University Press 1983.

[51] Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17**:368–376.

[52] Loytynoja A, Goldman N: **A model of evolution and structure for multiple sequence alignment.** *Philos Trans R Soc Lond B Biol Sci* 2008, **363**:3913–3919.

[53] Nielsen R, Yang Z: **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* 1998, **148**:929–936.

[54] Yang Z: **Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution.** *Mol Biol Evol* 1998, **15**:568–573.

[55] Yang Z, Nielsen R, Goldman N, Pedersen AM: **Codon-substitution models for heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155**:431–449.

[56] Yang Z: **Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A.** *J Mol Evol* 2000, **51**:423–432.

[57] Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985, **22**:160–174.

[58] Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC, Eaton M, Hamady M, Lindsay H, Liu Z, Lozupone C, McDonald D, Robeson M, Sammut R, Smit S, Wakefield MJ, Widmann J, Wikman S, Wilson S, Ying H, Huttley GA: **PyCogent: a toolkit for making sense from sequence.** *Genome Biol* 2007, **8**:R171.

[59] Lichtenauer-Kaligis EG, van der Velde-van Dijke I, den Dulk H, van de Putte P, Giphart-Gassler M, Tasseron-de Jong JG: **Genomic position in-**

fluences spontaneous mutagenesis of an integrated retroviral vector containing the HPRT cDNA as target for mutagenesis. *Hum Mol Genet* 1993, **2**:173–182.

[60] Horn PJ, Peterson CL: **Molecular biology. Chromatin higher order folding–wrapping up transcription.** *Science* 2002, **297**:1824–1827.

[61] Gontijo AMdMC, Green CM, Almouzni G: **Repairing DNA damage in chromatin.** *Biochimie* 2003, **85**:1133–1147.

[62] Verger A, Crossley M: **Chromatin modifiers in transcription and DNA repair.** *Cell Mol Life Sci* 2004, **61**:2154–2162.

[63] Widlak P, Pietrowska M, Lanuszewska J: **The role of chromatin proteins in DNA damage recognition and repair.** *Histochem Cell Biol* 2006, **125**:119–126.

[64] Jagannathan I, Cole HA, Hayes JJ: **Base excision repair in nucleosome substrates.** *Chromosome Res* 2006, **14**:27–37.

[65] Groth A, Rocha W, Verreault A, Almouzni G: **Chromatin challenges during DNA replication and repair.** *Cell* 2007, **128**:721–733.

[66] Hara R, Mo J, Sancar A: **DNA damage in the nucleosome core is refractory to repair by human excision nuclease.** *Mol Cell Biol* 2000, **20**:9173–9181.

[67] Chandley AC, Kofman-Alfaro S: **"Unscheduled" DNA synthesis in human germ cells following UV irradiation.** *Exp Cell Res* 1971, **69**:45–48.

[68] Sega GA: **Unscheduled DNA synthesis in the germ cells of male mice exposed in vivo to the chemical mutagen ethyl methanesulfonate.** *Proc Natl Acad Sci U S A* 1974, **71**:4955–4959.

[69] T B: **Nuclear envelope and chromatin structure.** *Int Rev Cytol Suppl* 1987, **17**:496–571.

[70] Balhorn R, Weston S, Thomas C, Wyrobek AJ: **DNA packaging in mouse spermatids. Synthesis of protamine variants and four transition proteins.** *Exp Cell Res* 1984, **150**:298–308.

[71] Boulikas T: **Evolutionary consequences of nonrandom damage and repair of chromatin domains.** *J Mol Evol* 1992, **35**:156–180.

[72] Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA: **8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G—-T and A—-C substitutions.** *J Biol Chem* 1992, **267**:166–172.

[73] Elango N, Kim SH, Vigoda E, Yi SV: **Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation.** *PLoS Comput Biol* 2008, **4**:e1000015.

[74] Misawa K, Kikuno RF: **Evaluation of the effect of CpG hypermutability on human codon substitution.** *Gene* 2009, **431**:18–22.

[75] Shiraishi M, Oates AJ, Sekiya T: **An overview of the analysis of DNA methylation in mammalian genomes.** *Biol Chem* 2002, **383**:893–906.

[76] Gross DS, Garrard WT: **Nuclease hypersensitive sites in chromatin.** *Annu Rev Biochem* 1988, **57**:159–197.

[77] Elgin SC: **The formation and function of DNase I hypersensitive sites in the process of gene activation.** *J Biol Chem* 1988, **263**:19259–19262.

[78] Jakobovits EB, Bratosin S, Aloni Y: **A nucleosome-free region in SV40 minichromosomes.** *Nature* 1980, **285**:263–265.

[79] Elgin SC: **DNAase I-hypersensitive sites of chromatin.** *Cell* 1981, **27**:413–415.

[80] Tuan D, London IM: **Mapping of DNase I-hypersensitive sites in the upstream DNA of human embryonic epsilon-globin gene in K562 leukemia cells.** *Proc Natl Acad Sci U S A* 1984, **81**:2718–2722.

[81] Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS: **DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays.** *Nat Methods* 2006, **3**:503–509.

[82] Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE: **High-resolution mapping and characterization of open chromatin across the genome.** *Cell* 2008, **132**:311–322.

[83] Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P: **Ensembl 2009.** *Nucleic Acids Res* 2009, **37**:D690–7.

[84] Karolchik D, Hinrichs AS, Kent WJ: **The UCSC Genome Browser.** *Curr Protoc Bioinformatics* 2007, **Chapter 1**:Unit 1.4.

[85] Lindsay H, Yap VB, Ying H, Huttley GA: **Pitfalls of the most commonly used models of context dependent substitution.** *Biol Direct* 2009, **4**:10.

[86] Holm S: **A simple sequentially rejective multiple test procedure.** *Scandinavian Journal of Statistics* 1979, **6**:65–70.

[87] Szent-Gyorgyi C, Finkelstein DB, Garrard WT: **Sharp boundaries demarcate the chromatin structure of a yeast heat-shock gene.** *J Mol Biol* 1987, **193**:71–80.

[88] Vyas P, Vickers MA, Simmons DL, Ayyub H, Craddock CF, Higgs DR: **Cis-acting sequences regulating expression of the human alpha-globin cluster lie within constitutively open chromatin.** *Cell* 1992, **69**:781–793.

[89] Wakeley J: **Substitution-rate variation among sites and the estimation of transition bias.** *Mol Biol Evol* 1994, **11**:436–442.

[90] Tsunoyama K, Bellgard MI, Gojobori T: **Intragenic variation of synonymous substitution rates is caused by nonrandom mutations at methylated CpG.** *J Mol Evol* 2001, **53**:456–464.

[91] Davey CS, Pennings S, Reilly C, Meehan RR, Allan J: **A determining influence for CpG dinucleotides on nucleosome positioning in vitro.** *Nucleic Acids Res* 2004, **32**:4322–4331.

[92] Holmquist GP: **Role of replication time in the control of tissue-specific gene expression.** *Am J Hum Genet* 1987, **40**:151–173.

[93] Leadon SA, Lawrence DA: **Strand-selective repair of DNA damage in the yeast GAL7 gene requires RNA polymerase II.** *J Biol Chem* 1992, **267**:23175–23182.

[94] Bedoyan J, Gupta R, Thoma F, Smerdon MJ: **Transcription, nucleosome stability, and DNA repair in a yeast minichromosome.** *J Biol Chem* 1992, **267**:5996–6005.

[95] Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SCJ, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison

RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CWH, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JNS, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PIW, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrimsdottir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VVB, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**:799–816.

[96] Widlak P, Garrard WT: **Nucleosomes and regulation of gene expression. Structure of the HIV-1 5'LTR.** *Acta Biochim Pol* 1998, **45**:209–219.

[97] Lodha M, Schroda M: **Analysis of chromatin structure in the control regions of the chlamydomonas HSP70A and RBCS2 genes.** *Plant Mol Biol* 2005, **59**:501–513.

[98] Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ: **Crystal structure of the nucleosome core particle at 2.8 A resolution.** *Nature* 1997, **389**:251–260.

[99] Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF: **Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome.** *Nature* 2007, **446**:572–576.

[100] Aboussekhra A, Biggerstaff M, Shivji MK, Vilpo JA, Moncollin V, Podust VN, Protic M, Hubscher U, Egly JM, Wood RD: **Mammalian DNA nucleotide excision repair reconstituted with purified protein components.** *Cell* 1995, **80**:859–868.

[101] Smerdon MJ, Thoma F: **Site-specific DNA repair at the nucleosome level in a yeast minichromosome.** *Cell* 1990, **61**:675–684.

[102] Suter B, Livingstone-Zatchej M, Thoma F: **Chromatin structure modulates DNA repair by photolyase in vivo.** *EMBO J* 1997, **16**:2150–2160.

[103] Wellinger RE, Thoma F: **Nucleosome structure and positioning modulate nucleotide excision repair in the non-transcribed strand of an active gene.** *EMBO J* 1997, **16**:5046–5056.

[104] Moggs JG, Almouzni G: **Chromatin rearrangements during nucleotide excision repair.** *Biochimie* 1999, **81**:45–52.

[105] Li S, Smerdon MJ: **Nucleosome structure and repair of N-methylpurines in the GAL1-10 genes of Saccharomyces cerevisiae.** *J Biol Chem* 2002, **277**:44651–44659.

[106] Smerdon MJ, Tlsty TD, Lieberman MW: **Distribution of ultraviolet-induced DNA repair synthesis in nuclease sensitive and resistant regions of human chromatin.** *Biochemistry* 1978, **17**:2377–2386.

[107] Bodell WJ: **Nonuniform distribution of DNA repair in chromatin after treatment with methyl methanesulfonate.** *Nucleic Acids Res* 1977, **4**:2619–2628.

[108] Gong F, Kwon Y, Smerdon MJ: **Nucleotide excision repair in chromatin and the right of entry.** *DNA Repair (Amst)* 2005, **4**:884–896.

[109] Suter B, Thoma F: **DNA-repair by photolyase reveals dynamic properties of nucleosome positioning in vivo.** *J Mol Biol* 2002, **319**:395–406.

[110] McGhee JD, Felsenfeld G: **Nucleosome structure.** *Annu Rev Biochem* 1980, **49**:1115–1156.

[111] Higasa K, Hayashi K: **Periodicity of SNP distribution around transcription start sites.** *BMC Genomics* 2006, **7**:66.

[112] Lee W, Tillo D, Bray N, Morse R, Davis R, Hughes T, Nislow C: **A high-resolution atlas of nucleosome occupancy in yeast.** *Nat Genet* 2007, **39**:1235–1244.

[113] Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek J, Costa G, McKernan K, Sidow A, Fire A, Johnson S: **A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning.** *Genome Res* 2008, **18**:1051–1063.

[114] Schmid CD, Bucher P: **ChIP-Seq data reveal nucleosome architecture of human promoters.** *Cell* 2007, **131**:831–832.

[115] Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K: **Dynamic regulation of nucleosome positioning in the human genome.** *Cell* 2008, **132**:887–898.

[116] Zhang Y, Shin H, Song JS, Lei Y, Liu XS: **Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq.** *BMC Genomics* 2008, **9**:537.

[117] Yang Z: **Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites.** *Mol Biol Evol* 1993, **10**:1396–1401.

[118] Wakeley J: **Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA.** *J Mol Evol* 1993, **37**:613–623.

[119] Yang Z: **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods.** *J Mol Evol* 1994, **39**:306–314.

[120] Durbin R: *Biological sequence analysis: probabalistic models and nucleic acids.* Cambridge, UK: Cambridge University Press 1998.

[121] Wakefield MJ, Maxwell P, Huttley GA: **Vestige: maximum likelihood phylogenetic footprinting.** *BMC Bioinformatics* 2005, **6**:130.

[122] Kunsch HR: **The jackknife and the bootstrap for general stationary observations**. *Annals of Statistics* 1989, **17**:1217–1241.

[123] Tretter S: **Estimating the frequency of a noisy sinusoid by linear regression**. *IEEE Trans Information Theory* 1985, **31**:832–835.

[124] Liang H, Lin YS, Li WH: **Fast evolution of core promoters in primate genomes.** *Mol Biol Evol* 2008, **25**:1239–1244.

[125] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129**:823–837.

[126] Daniel L Hartl AGC: *Principles of population genetics.* Sinauer Associates, Inc., 4th ed edition 2007.

[127] Sharkey M, Graba Y, Scott MP: **Hox genes in evolution: protein surfaces and paralog groups.** *Trends Genet* 1997, **13**:145–151.

[128] Merabet S, Hudry B, Saadaoui M, Graba Y: **Classification of sequence signatures: a guide to Hox protein function.** *Bioessays* 2009, **31**:500–511.

[129] Tornaletti S, Pfeifer GP: **Complete and tissue-independent methylation of CpG sites in the p53 gene: implications for mutations in human cancers.** *Oncogene* 1995, **10**:1493–1499.

[130] Rabinowicz PD, Palmer LE, May BP, Hemann MT, Lowe SW, McCombie WR, Martienssen RA: **Genes and transposons are differentially methylated in plants, but not in mammals.** *Genome Res* 2003, **13**:2658–2664.

[131] Jabbari K, Caccio S, Pais de Barros JP, Desgres J, Bernardi G: **Evolutionary changes in CpG and methylation levels in the genome of vertebrates.** *Gene* 1997, **205**:109–118.

[132] Karlin S, Campbell AM, Mrazek J: **Comparative DNA analysis across diverse genomes.** *Annu Rev Genet* 1998, **32**:185–225.

[133] Goldman N, Yang Z: **A codon-based model of nucleotide substitution for protein-coding DNA sequences.** *Mol Biol Evol* 1994, **11**:725–736.

[134] Muse SV, Gaut BS: **A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome.** *Mol Biol Evol* 1994, **11**:715–724.

[135] Huttley GA: **Modeling the impact of DNA methylation on the evolution of BRCA1 in mammals.** *Mol Biol Evol* 2004, **21**:1760–1768.

[136] Yap VB, Lindsay H, Esteal S, Huttley GA: **Estimates of the effect of natural selection on protein coding content**. *Mol Biol Evol* 2009, **In**

press.

[137] Schmidt S, Gerasimova A, Kondrashov FA, Adzhubei IA, Kondrashov AS, Sunyaev S: **Hypermutable non-synonymous sites are under stronger negative selection.** *PLoS Genet* 2008, **4**:e1000281.

[138] Matassi G, Sharp PM, Gautier C: **Chromosomal location effects on gene sequence evolution in mammals.** *Curr Biol* 1999, **9**:786–791.

[139] Smith NGC, Webster MT, Ellegren H: **Deterministic mutation rate variation in the human genome.** *Genome Res* 2002, **12**:1350–1356.

[140] Ellegren H, Smith NGC, Webster MT: **Mutation rate variation in the mammalian genome.** *Curr Opin Genet Dev* 2003, **13**:562–568.

[141] Lanave C, Preparata G, Saccone C, Serio G: **A new method for calculating evolutionary substitution rates.** *J Mol Evol* 1984, **20**:86–93.

[142] Pond SK, Muse SV: **Site-to-site variation of synonymous substitution rates.** *Mol Biol Evol* 2005, **22**:2375–2385.

[143] Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423**:241–254.

[144] Loytynoja A, Goldman N: **An algorithm for progressive multiple alignment of sequences with insertions.** *Proc Natl Acad Sci U S A* 2005, **102**:10557–10562.

[145] Antonarakis SE, McKusick VA: **OMIM passes the 1,000-disease-gene mark.** *Nat Genet* 2000, **25**:11.

[146] Proffitt JH, Davie JR, Swinton D, Hattman S: **5-Methylcytosine is not detectable in Saccharomyces cerevisiae DNA.** *Mol Cell Biol* 1984, **4**:985–988.

[147] Krawczak M, Ball EV, Cooper DN: **Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes.** *Am J Hum Genet* 1998, **63**:474–488.

[148] Johnson RE, Washington MT, Prakash S, Prakash L: **Fidelity of human DNA polymerase eta.** *J Biol Chem* 2000, **275**:7447–7450.

[149] Radford IR, Lobachevsky PN: **Clustered DNA lesion sites as a source of mutations during human colorectal tumourigenesis.** *Mutat Res* 2008, **646**:60–68.

[150] Bennetzen JL, Hall BD: **Codon selection in yeast.** *J Biol Chem* 1982, **257**:3026–3031.

[151] Schorderet DF, Gartler SM: **Analysis of CpG suppression in methylated and nonmethylated species.** *Proc Natl Acad Sci U S A* 1992, **89**:957–961.

[152] Kliman RM, Irving N, Santiago M: **Selection conflicts, gene expression, and codon usage trends in yeast.** *J Mol Evol* 2003, **57**:98–109.

[153] Jupe ER, Magill JM, Magill CW: **Stage-specific DNA methylation in a fungal plant pathogen.** *J Bacteriol* 1986, **165**:420–423.

[154] Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JPZ, Widom J: **A genomic code for nucleosome positioning.** *Nature* 2006, **442**:772–778.

[155] Segal MR: **Re-cracking the nucleosome positioning code.** *Stat Appl Genet Mol Biol* 2008, **7**:Article14.